UNIVERSITÄT KAISERSLAUTERN

Fachbereich Mathematik

Numerical Solution of a Nonstandard Darcy Flow Model

Master Thesis

Vsevolod Laptev

Supervisors
Prof. Dr. rer. nat. Helmut Neunzert
Dr. Aivars Zemitis

September 28, 1999

I hereby declare that I am the only author of this thesis and that no sources other than those listed have been used in this work.

Vsevolod Laptev

Kaiserslautern September 28, 1999

Acknowledgement

I would like to express my deep gratitude to my supervisors, Prof. Dr. H. Neunzert and Dr. A. Zemitis for giving me a very interesting mathematical problem and helping in my work on it.

I thank the Industrial Mathematics/Mathematics International program for the financial support, care, perfect organized study process, cultural programs.

I am very thankful to Prof. Dr. H. Neunzert and late Prof. A. V. Lukshin for giving me opportunity to study in Germany.

Contents

Abstract

We consider a Darcy flow model with saturation-pressure relation extended with a dynamic term, namely, the time derivative of the saturation. This model was proposed in works [1], [9], [10]. We restrict ourself to one spatial dimension and strictly positive initial saturation. For this case we transform the initial-boundary value problem into combination of elliptic boundary-value problem and initial value problem for abstract Ordinary Differential Equation. This splitting is rather helpful both for theoretical aspects and numerical methods.

1 Introduction

We consider two phase flows in porous media. They occur in various practical tasks, like unsaturated groundwater flow, oil recovery. The aim of mathematical approach here is to predict the saturation profiles if initial profiles and boundary conditions are known. For terminology, definitions and theory of flow through porous media we refer to [11],[12]. Basic notations were taken like in [1].

$S(t,x)$ is the level of saturation of a wetting phase ($S \in [0,1]$). We assume that wetting and nonwetting phases are water and air respectively. ϕ is the porosity of the porous medium. Typically the relation for the pressure difference in the phases is used

$$p_n - p_w = P(S) \tag{1}$$

where p_n is the pressure in the air (we assume that it has a constant value of atmospheric pressure) and p_w the pressure in the wetting phase. $P(S)$ is assumed to be a known bounded decreasing function on the water saturation, with $P(1) = 0$. The hydraulic conductivity $K(S)$ is a known nonnegative increasing function.

The differential equation describing dynamics of unsaturated flow is obtained by combining mass conservation of water

$$\phi \frac{\partial S}{\partial t} + \operatorname{div} \vec{q} = 0$$

with expression for the flux q from the Darcy's law

$$\vec{q} = K(S)(-\operatorname{grad} p_w + \rho \vec{g}) = K(S)(\operatorname{grad} P(S) + \rho \vec{g}) \tag{2}$$

where ρ is the water density and $\vec{g} = (-g, 0, 0)$. The result is the Richard's equation:

$$\phi \frac{\partial S}{\partial t} = -\operatorname{div}\left[K(S)(\operatorname{grad} P(S) + \rho \vec{g})\right] \tag{3}$$

In [1], [9], [10] the authors suggested a modified expression for (1):

$$p_n - p_w = P(S) - L\frac{\partial S}{\partial t}, \quad (L > 0) \tag{1'}$$

This lead to nonstandard Darcy flow model

$$\vec{q} = K(S)(-\operatorname{grad} p_w + \rho \vec{g}) = K(S)\left[\operatorname{grad}\left(P(S) - L\frac{\partial S}{\partial t}\right) + \rho \vec{g}\right] \tag{2'}$$

and modified Richard's equation:

$$\phi \frac{\partial S}{\partial t} = -\operatorname{div}\left[K(S)(\operatorname{grad} P(S) + \rho \vec{g})\right] + \operatorname{div}\left[K(S) L \operatorname{grad} \frac{\partial S}{\partial t}\right] \tag{3'}$$

We restrict ourself to the case of one spatial dimension where the main equation (3') becomes

$$\phi \frac{\partial S}{\partial t} = -\frac{\partial}{\partial x}\left[K(S)\left(\frac{\partial P(S)}{\partial x} - \rho g\right)\right] + \frac{\partial}{\partial x}\left[K(S) L \frac{\partial^2 S}{\partial x \partial t}\right] \tag{4}$$

$S = S(t,x)$ - saturation, $t \in \Pi = (0,T)$, $x \in \Omega = (0,l)$.
Let

$$F(S) = K(S)\left(\frac{\partial P(S)}{\partial x} - \rho g\right) = K(S)\left(P'(S)\frac{\partial S}{\partial x} - \rho g\right) \tag{5}$$

To complete the problem (4) we need to add initial and boundary conditions.
Initial conditions: $S(0,x)$ is given.
Boundary conditions. We will deal with two types of B.C.
B.C.1: a flux is zero on boundaries $K(S)L\frac{\partial^2 S}{\partial x \partial t} - F(S) = 0$, $x = 0, l$.
B.C.2: $S(t,0)$, $S(t,l)$ are given for all $t \in [0,T]$. We will use only the simplest case: constant values of S on the boundary.

2 General Aspects

2.1 Another formulation of problem (4)

It is convenient to rename $u(t,x) = \frac{\partial S}{\partial t}(t,x)$. Another form of equation (4) with variables u and S:

$$-\frac{\partial}{\partial x}\left[K(S)L\frac{\partial u}{\partial x}\right] + \phi u = -\frac{\partial}{\partial x}F(S) \tag{6}$$

With boundary conditions:
B.C.1 : $K(S)L\frac{\partial u}{\partial x} - F(S) = 0,\ x = 0, l$
B.C.2 : $S(t,0) = S_l, \quad S(t,l) = S_r, \quad u(t,0) = 0, \quad u(t,l) = 0.$

Assume that S is known for some t. Then we have an elliptic equation on $u(t,\cdot)$: $L_1 u + u = -L_2 S$ where $L_1 u$, $L_1 u + u$ are elliptic. It is possible to use well developed theory of second order elliptic equations to get existence and some properties of solution operator acting on S. If $S(x) > 0$ for all x then operators are strictly elliptic ($K(0) = 0$ - no ellipticity).

For given $S > 0$ equation (6) is elliptic with respect to u and we can use existence, uniqueness results for elliptic equations

Weak solution approach: multiplying (6) by test function $v(x)$ and integrating by parts on Ω:

$$\int\limits_\Omega K(S)L\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}\,dx - \int\limits_{\partial\Omega} K(S)L\frac{\partial u}{\partial x}v\,d\sigma + \int\limits_\Omega \phi uv\,dx = \int\limits_\Omega F(S)\frac{\partial v}{\partial x}\,dx - \int\limits_{\partial\Omega} F(S)v\,d\sigma \tag{7}$$

Boundary condition 1: Zero flux means that boundary integrals together are zero. We can use functional space $H^1(\Omega) = W^{1,2}(\Omega)$.

Boundary condition 2: Constant values on the boundary $S(t,0) = S_l$, $S(t,l) = S_r$, $u(t,0) = 0$, $u(t,l) = 0$. We can use functional space $H_0^1(\Omega) = W_0^{1,2}(\Omega)$ for Dirichlet problem.

Let $V = H^1(\Omega)$ for B.C.1 and $V = H_0^1(\Omega)$ for B.C.2. Weak formulation for the elliptic problem (6): Find $u \in V$ such that for any $v \in V$

$$a(u,v) + (u,v)_{\phi,0} = l(v) \tag{8}$$

where $a(u,v) = \int\limits_\Omega K(S)L\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}\,dx$, $(u,v)_{\phi,0} = \int\limits_\Omega \phi uv\,dx$, $l(v) = \int\limits_\Omega F(S)\frac{\partial v}{\partial x}\,dx$

Assumptions

$$\begin{array}{ll} 1. & S(x) \overset{a.e}{\geq} S_* > 0 \Rightarrow K(S) \geq K(S_*) = \mu[S] > 0; \\ 2. & \phi \in L^\infty(\Omega), 1 \geq \phi \geq \phi_0 > 0 \\ 3. & K, KP' - \text{ Lipschitz continuous, bounded :} \end{array} \tag{9}$$

$$|K(S_1) - K(S_2)| \leq L_K|S_1 - S_2|, \quad |K(S_1)P'(S_1) - K(S_2)P'(S_2)| \leq L_{KP'}|S_1 - S_2|.$$

$a(u,v) + (u,v)_{\phi,0}$ is a symmetric bilinear functional. We can check V-ellipticity and continuity of this functional:

$$(E) \quad a(u,u) + (u,u)_{\phi,0} \geq \mu[S]L\left\|\frac{\partial u}{\partial x}\right\|_0^2 + \phi_0\|u\|_0^2 \geq \min\{\mu[S]L, \phi_0\}\|u\|_1^2 = E[S]\|u\|_1^2$$

where $E[S] > 0$ - ellipticity constant.

$$(C) \quad |a(u,v) + (u,v)_{\phi,0}| \leq \|K\|_\infty L\left\|\frac{\partial u}{\partial x}\right\|_0\left\|\frac{\partial v}{\partial x}\right\|_0 + \|u\|_0\|v\|_0 \leq C\|u\|_1\|v\|_1$$

where $C > 0$ - continuity constant.

$l = l[S] \in V'$ - adjoint space for V. $l(v)$ is a linear bounded functional on V.

$$|l[S](v)| = \left| \int_\Omega K(S)P'(S)\frac{\partial S}{\partial x}\frac{\partial v}{\partial x}\,dx - \int_\Omega K(S)\rho g\frac{\partial v}{\partial x}\,dx \right| \leq$$

$$\leq \|KP'\|_\infty \left\|\frac{\partial S}{\partial x}\right\|_0 \left\|\frac{\partial v}{\partial x}\right\|_0 + \rho g\|K\|_\infty \mu(\Omega) \left\|\frac{\partial v}{\partial x}\right\|_0 \leq$$

$$\leq \left(\|KP'\|_\infty \|S\|_1 + \rho g\|K\|_\infty \mu(\Omega)\right)\|v\|_1$$

$$\|l[S]\|_{V'} \leq \|KP'\|_\infty \|S\|_1 + \rho g\|K\|_\infty \mu(\Omega) \tag{10}$$

From [3, Lemma 3.18, part a) p. 97] we can get existence-uniqueness results: The problem (8) has unique solution $u \in V$ and

$$\|u\|_V \leq \frac{1}{E[S]}\|l[S]\|_{V'} = \frac{1}{E[S]} \sup_{v\neq 0, v\in V} \frac{|l[S](v)|}{\|v\|_1}.$$

We can introduce a solution operator $A : H^1(\Omega) \to V$.
$u = A(S)$ - a unique solution of problem (8) corresponding to S. In terms of A, (4) can be written in a form:

$$u = \frac{dS}{dt} = A(S), \qquad S(0) = S_0, \qquad S(t) \in H^1(\Omega) \quad \forall t \in \Pi \tag{11}$$

We don't know exactly the domain of definition for A, but at least it contains all functions $S \in H^1$ that are bounded away from zero: $\exists$ constant $S_* > 0$ that $S(x) \geq S_*$ almost everywhere.

Let $U_b = \{S \in H^1(\Omega) : \|S - S_0\|_1 \leq b\}$ a neighborhood of S_0.

In one dimensional case we can use embedding $H^1(\Omega)$ to the space of continuous bounded functions $C_B(\Omega)$ with supremum norm (see [2, p. 97]):
a)If Ω has cone property, $mp = 2 > n = 1$ then $W^{m,p} = W^{1,2}(\Omega) \to C_B(\Omega)$
b)If Ω has strong local Lipschitz property, $mp = 2 > n = 1 > (m-1)p = 0$ then $W^{m,p} = W^{1,2}(\Omega) \to C^{0,\lambda}(\bar{\Omega})$, $0 < \lambda \leq m - (n/p)$ (for example $\lambda = 1/2$)
In one dimension case see [5, p. 31]

From these results we need: $u \in H^1(\Omega) \Rightarrow u \in C(\Omega)$ and $\operatorname{ess\,sup}_{x\in\Omega} |u| \leq C^B\|u\|_1$

If the initial value is bounded away from zero: $S_0(x) \geq S_{0*} > 0$ a.e. then there exist b and $S_* > 0$ that: $\forall S \in U_b$, $S \geq S_*$ a.e.; in other words, U_b is bounded away from zero.

Boundedness of U_b: $\|S\|_1 \leq \|S_0\|_1 + b$.

Also there exist constants μ, E that $\mu[S] \geq \mu$, $E[S] \geq E$.

Remark In our case S is saturation and we also need to have $S \leq 1$. And if $S_0 < 1$ then at the same way we can choose ball U_b bounded away from 1. (but we will not mention it explicitly).

2.2 Properties of $A(S)$

A1. $A(S)$ **is bounded on** U_b.

From (10): $$\|A(S)\|_1 = \|u\|_1 \leq \frac{1}{E}\left(\|KP'\|_\infty(\|S_0\|_1 + b) + \rho g\|K\|_\infty\mu(\Omega)\right) = B$$

For any $S \in U_b$: $\|A(S)\|_1 \le B$.

A2. $A(S)$ **is Lipschitz continuous on** U_b:

$$\|A(S_1) - A(S_2)\|_1 \le L_A \|S_1 - S_2\|_1 \qquad \forall S_1, S_2 \in U_b.$$

$u_1 = A(S_1)$, $u_2 = A(S_2)$. $a(u,v)$, $l(v)$ depend on S,

$$a[S](u,v) = \int_\Omega K(S) L \frac{\partial u}{\partial x}\frac{\partial v}{\partial x}\,dx, \qquad l[S](v) = \left(F(S), \frac{\partial v}{\partial x}\right)_0.$$

$$\begin{array}{l} a[S_1](u_1,v) + (u_1,v)_{\phi,0} = l[S_1](v), \\ a[S_2](u_2,v) + (u_2,v)_{\phi,0} = l[S_2](v) \end{array} \qquad \forall v \in V.$$

$$a[S_1](u_2,v) + (u_2,v)_{\phi,0} = \int_\Omega K(S_1) L \frac{\partial u_2}{\partial x}\frac{\partial v}{\partial x}\,dx + \int_\Omega \phi u_2 v\,dx =$$

$$= l[S_2](v) - \int_\Omega [K(S_2) - K(S_1)] L \frac{\partial u_2}{\partial x}\frac{\partial v}{\partial x}\,dx$$

Substitute $a[S_1](u_1,v) + (u_1,v)_{\phi,0} = l[S_1](v)$

$$a[S_1](u_2-u_1,v) + (u_2-u_1,v)_{\phi,0} = l[S_2](v) - l[S_1](v) + \int_\Omega [K(S_1) - K(S_2)] L \frac{\partial u_2}{\partial x}\frac{\partial v}{\partial x}\,dx$$

$$= l[S_2](v) - l[S_1](v) + l[S_1, S_2, u_2](v) = \mathcal{L}(v) \text{for any } v \in V.$$

So $u_2 - u_1$ is a solution of $a[S_1](u_2-u_1,v) + (u_2-u_1,v)_{\phi,0} = \mathcal{L}(v)$, where $\mathcal{L}(v)$ is a linear bounded functional on V.

$$\|A(S_2) - A(S_1)\|_1 = \|u_2 - u_1\|_1 \le \|\mathcal{L}\|_{V'}/E \tag{12}$$

and we have to estimate $\|\mathcal{L}\|_{V'} = \sup\limits_{v \in V, \|v\|_V = 1} |\mathcal{L}(v)|$.

$$\|\mathcal{L}\|_{V'} \le \|l[S_2] - l[S_1]\|_{V'} + \|l[S_1, S_2, u_2]\|_{V'} \tag{13}$$

a) estimation for $l[S_1, S_2, u_2](v) = \int\limits_\Omega [K(S_1) - K(S_2)] L \frac{\partial u_2}{\partial x}\frac{\partial v}{\partial x}\,dx$:

- $|K(S_1(x)) - K(S_2(x))| \le L_K |S_1(x) - S_2(x)| \overset{a.e}{\le} L_K C^B \|S_1 - S_2\|_1 = C_1^B \|S_1 - S_2\|_1$

- embedding to $C_B(\Omega)$.

$$|l[S_1, S_2, u_2](v)| \le C_1^B L \|S_1 - S_2\|_1 \|\frac{\partial u_2}{\partial x}\|_0 \|\frac{\partial v}{\partial x}\|_0 \le C_1^B L \|S_1 - S_2\|_1 \|u_2\|_1 \|v\|_1$$

$$\|u_2\| \le B \qquad \Rightarrow \|l[S_1, S_2, u_2]\|_{V'} \le C_1^B L B \|S_1 - S_2\|_1 \tag{14}$$

- b) estimation for $l[S_2] - l[S_2]$:

$$|l[S_2](v) - l[S_1](v)| = |\langle F(S_2), \frac{\partial v}{\partial x}\rangle_0 - \langle F(S_1), \frac{\partial v}{\partial x}\rangle_0| =$$

$$= \left| \int\limits_\Omega K(S_2) \left(P'(S_2) \frac{\partial S_2}{\partial x} - \rho g \right) \frac{\partial v}{\partial x}\, dx - \int\limits_\Omega K(S_1) \left(P'(S_1) \frac{\partial S_1}{\partial x} - \rho g \right) \frac{\partial v}{\partial x}\, dx \right| \leq$$

$$\leq \left| \int\limits_\Omega \left(K(S_2) P'(S_2) \frac{\partial S_2}{\partial x} \frac{\partial v}{\partial x} - K(S_1) P'(S_1) \frac{\partial S_1}{\partial x} \frac{\partial v}{\partial x} \right) dx \right| +$$

$$+ \left| \int\limits_\Omega \rho g \left(K(S_1) - K(S_2) \right) \frac{\partial v}{\partial x}\, dx \right| = |I_1| + |I_2|$$

Let first integral be I_1, second - I_2. Next we will use embedding H^1 to C_B

$$|I_2| \leq \rho g C_1^B \|S_1 - S_2\|_1 \mu(\Omega) \|v\|_1$$

$$|I_1| \leq \left| \int\limits_\Omega \left[K(S_2) P'(S_2) - K(S_1) P'(S_1) \right] \frac{\partial S_2}{\partial x} \frac{\partial v}{\partial x}\, dx \right| +$$

$$+ \left| \int\limits_\Omega K(S_1) P'(S_1) \left(\frac{\partial S_2}{\partial x} - \frac{\partial S_1}{\partial x} \right) \frac{\partial v}{\partial x}\, dx \right| \leq$$

$|KP'(S_1) - KP'(S_2)| \leq L_{KP'} |S_1 - S_2|, \qquad C_2^B = L_{KP'} C^B$

$$\leq C_2^B \|S_2 - S_1\|_1 \|S_2\|_1 \|v\|_1 + \|KP'\|_\infty \|S_2 - S_1\|_1 \|v\|_1$$

$$\leq \left(C_2^B \left(\|S^0\|_1 + b \right) + \|KP'\|_\infty \right) \|S_2 - S_1\|_1 \|v\|_1.$$

$$\|l[S_2] - l[S_1]\|_{V'} \leq \left(C_2^B \left(\|S^0\|_1 + b \right) + \|KP'\|_\infty + \rho g C_1^B \mu(\Omega) \right) \|S_1 - S_2\|_1 \qquad (15)$$

• Altogether (12),(13),(14), (15) give:

$$\|A(S_2) - A(S_1)\|_1 \leq \left(C_2^B \left(\|S^0\|_1 + b \right) + \|KP'\|_\infty + \rho g C_1^B \mu(\Omega) + C_1^B L B \right) \|S_2 - S_1\|_1 / E$$

$$\|A(S_2) - A(S_1)\|_1 \leq L_A \|S_2 - S_1\|_1$$

$\Rightarrow A(S)$ - Lipschitz continuous on U_b •

A3. $A(S)$ **conserves smoothness**

We have already shown that for $S \in H^1$, $A(S) \in H^1$. It seems that this is also (formal) true for other classes of smooth functions, for example H^k, C^k (It is really, not formally true when the corresponding regularity theorems are valid). We can compare $A(S)$ with $-\frac{1}{\phi} \frac{\partial}{\partial x} F(S)$ from Richard's equation:

$$\frac{\partial S}{\partial t} = -\frac{1}{\phi} \frac{\partial}{\partial x} F(S).$$

$\frac{\partial}{\partial x} F(S)$ decreases smoothness of S, for example C^k to C^{k-2}, H^k to H^{k-2}. On the other hand $\frac{\partial}{\partial x} F(S)$ can be explicitly calculated from (5) and $A(S)$ can be obtained only by solving an elliptic equation.

2.3 On the well-posedness of (11)

Now we can compare (11) with Ordinary Differential Equation: $y' = f(t, y(t))$, $y(0) = y_0$. To have local existence and uniqueness of the solution of ODE we need boundedness and Lipschitz continuity of $f(t, y)$on y in some domain D around initial point $(0, y_0)$ (for instance $D = \{(t, y) : |y - y_0| < b,\ |t| < T\}$). $|f(t, y)| \le B$ in D, $|f(t, y_1) - f(t, y_2)| \le L|y_1 - y_2|$ for $(t, y_1), (t, y_2) \in D$. To prove the existence of ODE one can construct a sequence $y_k(t)$: $y_0(t) = y_0$,

$$y_{k+1}(t) = y_0 + \int_0^t f(\tau, y_k(\tau))\, d\tau,$$

that stays in D for t from some interval $[0, T]$ and converges to solution $y(t)$ that satisfies the integral form of ODE:

$$y(t) = y_0 + \int_0^t f(\tau, y(\tau))\, d\tau.$$

To prove uniqueness, the Gronwall's lemma can be used.

In our case properties **A1,A2** are similar to corresponding properties of $f(t, y)$ in ODE case. We can follow the ODE existence proof trying to fit it with our case. So let consider a sequence: $S_0(t) = S_0$,

$$S_{k+1}(t) = S_0 + \int_0^t A(S_k(\tau))\, d\tau \tag{16}$$

The integral in (16) we will consider as Riemann integral in a Banach space E. see [4, §38 pp 304-306, §39] for details. Some properties from [4] that we will explicitly use:

Def 1 Let $y(t) \in E\ \forall t \in [0, T]$. $y(t)$ is called continuous in $[0, T]$ if $\forall t \in [0, T]$: $\|y(t + h) - y(t)\|_E \to 0$ when $h \to 0$. Notation: $y \in C([0, T] \to E)$.

Def 2 Derivative of $y(t)$ at the point t:

$$\frac{d}{dt}y(t) = \lim_{h \to 0} \frac{1}{h}[y(t + h) - y(t)]$$

if the limit exists in the sense of E.

Y1. If $y(t)$ has a derivative at the point t, then $y(t)$ is continuous at the point t.

Y2. If $y(t)$ is Lipschitz continuous in $[0, T]$ (L.C.) ($\exists$ constant $L_y > 0$ such that $\forall t_1, t_2 \in [0, T]$: $\|y(t_1) - y(t_2)\|_E \le L_y|t_1 - t_2|$), then $y(t)$ is continuous in $[0, T]$.

Y3. If $y(t)$ is continuous in $[0, T]$, then $\|y(t)\|_E$ is a continuous real function from $C([0, T] \to R)$ and $\int_0^t \|y(\tau)\|_E\, d\tau$ is well defined.

Y4. If $y(t)$ is continuous in $[0, T]$, then the integral $\int_0^t y(\tau)\, d\tau$ is well defined in E. Moreover:

$$\left\| \int_0^t y(\tau)\, d\tau \right\|_E \le \int_0^t \|y(\tau)\|_E\, d\tau \tag{17}$$

Y5. If $y(t)$ is continuous in $[0, T]$, $c \in (0, T)$, then

$$\int_0^T y(\tau)\, d\tau = \int_0^c y(\tau)\, d\tau + \int_c^T y(\tau)\, d\tau.$$

Y6. If $y(t)$ is continuous in $[0, T]$, then the function $Y(t) = \int_0^t y(\tau)\, d\tau$ is differentiable in $[0, T]$, and thus $\frac{d}{dt}Y(t) = y(t)$.

Y7. If the function $Y(t)$ possesses a continuous derivative with respect to t, $\frac{d}{dt}Y(t) = y(t)$, then

$$\int_0^t y(\tau)\,d\tau = Y(t) - Y(0).$$

$E = H^1(\Omega)$ and $\|\cdot\|_E = \|\cdot\|_1$ when the opposite is not explicitly mentioned.

2.3.1 On local existence

Our goal now is to show that for some $T > 0$ (11) has unique solution in $C([0,T] \to H^1)$ that continuously depends on initial data in $C([0,T] \to H^1)$ provided S_0 is bounded away from zero. Our plan is to show that sequence $\{S_n\}_{n=0}^{\infty}$ from (16) is well defined (Steps 1-3) and converges (Step 4) to $S(t)$ - a solution of

$$S(t) = S_0 + \int_0^t A(S(\tau))\,d\tau, \tag{18}$$

(Step 5) and this implies that $S(t)$ is also a solution of (11) (Step 6).
Remark Here we consider $\frac{dS}{dt}$ is in the sense of **Def 2** in (11).

When $S_0(x) \geq S_{0*} > 0$ then we can choose appropriate U_b with constants b, B, L_A. Now we can determine $T > 0$:

$$T < \min\{1/L_A, b/B\} \tag{19}$$

For each element of sequence $\{S_n\}_{n=0}^{\infty}$ from (16) we have to show
a) That $S_n(t)$ is well defined element in H^1 for all $t \in [0,T]$.
b) For all $t \in [0,T]$, $S_n(t)$ stays in $U_b \subset H^1$.
c) $S_n(t)$ is Lipschitz continuous in $[0,T]$ with constant B.

We will use induction in three steps: Step 1 for $S_0(t)$, Step 2 for $S_1(t)$ - induction's base and Step 3 induction's hypothesis from $S_n(t)$ to $S_{n+1}(t)$ (step 3 is similar to step 2).
Step 1. $S_0(t) = S_0\ \forall t \in [0,T]$.
a) $S_0(t) \in H^1, \quad \forall t \in [0,T]$.
b) $S_0(t) \in U_b, \quad \forall t \in [0,T]$.
c) $S_0(t)$ has (L.C.) property with constant B: $\forall t_1, t_2 \in [0,T]$

$$\|S_0(t_1) - S_0(t_2)\|_1 = \|0\|_1 \leq B|t_1 - t_2|$$

Step 2. $S_1(t) = S_0 + \int_0^t A(S_0(\tau))\,d\tau$
a) From **A2**, properties b) and c) for $S_0(t)$ we can get $\forall t_1, t_2 \in [0,T]$:

$$\|A(S_0(t_1)) - A(S_0(t_2))\|_1 \leq L_A\|S_0(t_1) - S_0(t_2)\|_1 \leq L_A B|t_1 - t_2|.$$

It means that $A(S_0(t))$ has (L.C.) property, hence is integrable and $S_1(t) \in H^1$, $\forall t \in [0,T]$.
b) Using **Y4** for $A(S_0(t))$, b) for $S_0(t))$, **A1**

$$\|S_1(t) - S_0\|_1 = \left\| \int_0^t A(S_0(\tau))\,d\tau \right\|_1 \leq \int_0^t \|A(S_0(\tau))\|_1\,d\tau \leq Bt \leq BT < b$$

or $S_1(t) \in U_b$: $\forall t \in [0,T]$.
c) To obtain (L.C.) property for $S_1(t)$ we use **Y5**, **Y4** for $A(S_0(t))$, **A1**: $\forall t_1, t_2 \in [0,T]$

$$\|S_1(t_1) - S_1(t_2)\|_1 = \left\| \int_{t_2}^{t_1} A(S_0(\tau))\,d\tau \right\|_1 \leq \int_{t_2}^{t_1} \|A(S_0(\tau))\|_1\,d\tau \leq B|t_1 - t_2|$$

Step 3. Suppose that for S_n a),b),c) are valid:
a) $S_n(t) \in H^1$, $\forall t \in [0,T]$
b) $S_n(t) \in U_b$, $\forall t \in [0,T]$
c) (L.C) $\|S_n(t_1) - S_n(t_2)\|_1 \leq B|t_1 - t_2|$, $\forall t_1, t_2 \in [0,T]$
We need to show that a),b),c) are also valid for $S_{n+1}(t) = S_0 + \int_0^t A(S_n(\tau))\, d\tau$
To do this we can apply the same arguments like in **Step 2**:
a)

$$\|A(S_n(t_1)) - A(S_n(t_2))\|_1 \leq L_A \|S_n(t_1) - S_n(t_2)\|_1 \leq L_A B|t_1 - t_2|$$

$A(S_n(t))$ is (L.C)$\Rightarrow$ integrable $\Rightarrow$ $S_{n+1}(t) \in H^1$, $\forall t \in [0,T]$.
b) $S_{n+1}(t) \in U_b$, $\forall t \in [0,T]$ since

$$\|S_{n+1}(t) - S_0\|_1 = \left\| \int_0^t A(S_n(\tau))\, d\tau \right\|_1 \leq \int_0^t \|A(S_n(\tau))\|_1\, d\tau \leq Bt \leq BT < b$$

c) $\forall t_1, t_2 \in [0,T]$

$$\|S_{n+1}(t_1) - S_{n+1}(t_2)\|_1 = \left\| \int_{t_2}^{t_1} A(S_n(\tau))\, d\tau \right\|_1 \leq \int_{t_2}^{t_1} \|A(S_n(\tau))\|_1\, d\tau \leq B|t_1 - t_2|.$$

So we can define a sequence $\{S_n\}_{n=0}^{\infty}$ in $H^1(\Omega)$.
Step 4. Now we will investigate a convergence of this sequence.

$$S_{n+1}(t) = S_0 + \sum_{k=0}^{n} (S_{k+1}(t) - S_k(t))$$

$$\|S_{k+1}(t) - S_k(t)\|_1 = \left\| \int_0^t A(S_k(\tau))\, d\tau - \int_0^t A(S_{k-1}(\tau))\, d\tau \right\|_1 =$$

$$= \left\| \int_0^t [A(S_k(\tau)) - A(S_{k-1}(\tau))]\, d\tau \right\|_1 \leq$$

Function $[A(S_k(t)) - A(S_{k-1}(t))]$ is Lipschitz continuous on $[0,T]$ as a sum of Lipschitz continuous functions and we can use **Y4**. $[S_k(t) - S_{k-1}(t)]$ is also (L.C.). From **Y3**: $\|S_k(t) - S_{k-1}(t)\|_1$ approaches it's maximal value on $[0,T]$.

$$\leq \int_0^t \|A(S_k(\tau)) - A(S_{k-1}(\tau))\|_1\, d\tau \leq \int_0^t L_A \|S_k(\tau) - S_{k-1}(\tau)\|_1\, d\tau \leq$$

$$\leq L_A t \max_{t\in[0,T]} \|S_k(t) - S_{k-1}(t)\|_1 \leq L_A T \max_{t\in[0,T]} \|S_k(t) - S_{k-1}(t)\|_1 \leq \ldots$$

$$\ldots \leq (L_A T)^k \max_{t\in[0,T]} \|S_1(t) - S_0(t)\|_1 \leq (L_A T)^k b$$

from (19): $L_A T < 1$

$$\sum_{k=0}^{\infty} \|S_{k+1}(t) - S_k(t)\|_1 \leq \sum_{k=0}^{\infty} (L_A T)^k b < \infty$$

We have shown that $S_n(t)$ converges in $H^1(\Omega)$ uniformly on $[0,T]$ to some function $S(t)$. Let us show that $S(t)$ also has properties a), b), c).
a) $\forall t$: $\lim_{n\to\infty} S_n(t) = S(t)$ in $H^1(\Omega)$, uniformly on $t \in [0,T]$.

b) U_b - closed $\Rightarrow S(t) \in U_b$, $\forall t \in [0,T]$
c) $\circ$

$$\|S(t_1) - S(t_2)\|_1 \leq \|S(t_1) - S_n(t_1) + S_n(t_1) - S_n(t_2) + S_n(t_2) - S(t_2)\|_1 \leq$$

For any ϵ we can find N that for $n > N$, $\forall t \in [0,T]$ $\|S(t) - S_n(t)\|_1 < \varepsilon$

$$\leq \|S(t_1) - S_n(t_1)\|_1 + \|S_n(t_1) - S_n(t_2)\|_1 + \|S_n(t_2) - S(t_2)\|_1 \leq B|t_1 - t_2| + 2\varepsilon$$

ε can be chosen arbitrary small $\Rightarrow \|S(t_1) - S(t_2)\|_1 \leq B|t_1 - t_2|$ $\forall t \in [0,T]$. •
Step 5. Our task is to verify (18)
$\circ$ First: $A(S(t))$ is (L.C.), since $\forall t_1, t_2 \in [0,T]$

$$\|A(S(t_1)) - A(S(t_2))\|_1 \leq L_A \|S(t_1) - S(t_2)\|_1 \leq L_A B|t_1 - t_2|$$

and we can integrate $A(S(t))$ on t. Moreover $\int_0^t A(S_n(\tau))\,d\tau$ converges to $\int_0^t A(S(\tau))\,d\tau$ in $H^1(\Omega)$ uniformly on $t \in [0,T]$:

$$\left\| \int_0^t A(S_n(\tau))\,d\tau - \int_0^t A(S(\tau))\,d\tau \right\|_1 = \left\| \int_0^t [A(S_n(\tau)) - A(S(\tau))]\,d\tau \right\|_1 \leq$$

$[A(S_n(t)) - A(S(t))]$ – (L.C.) on t as a sum of (L.C.) functions; $S_n(t) \to S(t)$ when $n \to \infty$, uniformly on $t \in [0,T]$.

$$\leq \int_0^t \|A(S_n(\tau)) - A(S(\tau))\|_1\,d\tau \leq L_A \int_0^t \|S_n(\tau) - S(\tau)\|_1\,d\tau \leq L_A T \varepsilon_n \to 0, \quad n \to \infty$$

$$\Rightarrow \int_0^t A(S_n(\tau))\,d\tau \longrightarrow \int_0^t A(S(\tau))\,d\tau, \quad n \to \infty$$

$$S(t) \longleftarrow S_n(t) = S_0 + \int_0^t A(S_{n-1}(\tau))\,d\tau \longrightarrow S_0 + \int_0^t A(S(\tau))\,d\tau, \qquad n \to \infty$$

$\forall t \in [0,T]$, uniformly in t And we have (18). •
Step 6. $A(S(t))$ is continuous on $[0,T]$. From (18) and **Y6** we have (11)
We can summarize **Steps 1-6**: If the initial data is bounded away from zero then U_b can be chosen. It determines the constants b, B, L_A, and T from (19). Then for $t \in [0,T]$ one can construct $S(t)$ a solution of (11).
Remark We always had initial data at $t = 0$ and time interval $[0,T]$. It was not a restriction and if the solution is known at some time moment t_0, then we can consider $S(t_0) = S_0$ as initial data and look for solution at $[t_0, t_0 + T]$.

2.3.2 On local uniqueness, continuous dependence on initial data

Suppose $\tilde{S}(t)$ is continuous and satisfies (18) at $t \in [0,\tilde{T}]$. And S_0 is bounded away from zero. Then we can choose neighborhood U_b with constants b, B, L_A. For this U_b we can find T from (19).
U1. The function $\tilde{S}(t)$ stays inside U_b while $t \in [0, \min\{T, \tilde{T}\}]$.
$\circ$ Assume it is not true and there is $t_* \leq T, \tilde{T}$ that $\|\tilde{S}(t_*) - S_0\|_1 = b$ and $\|\tilde{S}(t) - S_0\|_1 < b$ for $t \in [0, t_*)$ (we can find t_* because $\|\tilde{S}(t) - S_0\|_1$ is continuous). Then for all $t \in [0, t_*]$ using that $S(t) \in U_b$ and (19) we can estimate:

$$\|\tilde{S}(t_*) - S_0\|_1 \leq \int_0^{t_*} \|A(\tilde{S}(\tau))\|_1\,d\tau \leq B t_* \leq BT < b$$

And we have contradiction with definition of t_*. •
Now assume that we have two solutions $\tilde{S}(t)$ and $S(t)$ of (11) at $t \in [0, \tilde{T}]$, with $\tilde{S}(t_0) = S(t_0) = S_0$ - bounded away from zero. We can choose U_b around $S_0 > 0$ with constants b, B, L_A; T from (19) but not greater than $\tilde{T}$. From **U1**, $\tilde{S}(t)$, $S(t)$ stays inside U_b while $t \in [t_0, t_0 + T]$. Using that $\frac{d\tilde{S}}{dt}$, $\frac{dS}{dt}$ are in the sense of **Def 2**, from **Y1** we can conclude that $\tilde{S}(t), S(t)$ are continuous in $[0, T]$. Then $A(\tilde{S}(t)), A(S(t))$ are continuous in $[0, T]$. From **Y7** we conclude that $\tilde{S}(t)$, $S(t)$ satisfy (18) at $[0, T]$. Then

$$S(t) - \tilde{S}(t) = \int_{t_0}^{t} [A(S(\tau)) - A(\tilde{S}(\tau))]\, d\tau$$

$[A(S(t)) - A(\tilde{S}(t))]$ is a continuous function on $[0, T]$. From **Y4**, **A2** we obtain an integral estimation:

$$\|S(t) - \tilde{S}(t)\|_1 \leq \int_{t_0}^{t} \|A(S(\tau)) - A(\tilde{S}(\tau))\|_1\, d\tau \leq L_A \int_{t_0}^{t} \|S(\tau) - \tilde{S}(\tau)\|_1\, d\tau$$

Lemma(Gronwall) (see[6], p.5) Assume that for $t \in [t_0, t_0 + a]$

$$\phi(t) \leq \delta_1 + \delta_2 \int_{t_0}^{t} \psi(\tau)\phi(\tau)\, d\tau$$

where $\phi(t), \psi(t) \geq 0$ are continuous at $[t_0, t_0 + a]$, $\delta_1, \delta_2 > 0$. Then for $t \in [t_0, t_0 + a]$:

$$\phi(t) \leq \delta_1 \exp\left(\delta_2 \int_{t_0}^{t} \psi(\tau)\, d\tau\right).$$

Corollary. If $\delta_1 = 0$, then $\phi(t) = 0$ at $[t_0, t_0 + a]$.
In our case $\phi(t) = \|S(t) - \tilde{S}(t)\|_1$ - continuous (from **Y3**), $\psi(t) = 1$, $\delta_2 = L_A$, $\delta_1 = 0$, $a = T$. From the Corollary we conclude that $S(t) = \tilde{S}(t)$ at $t \in [t_0, t_0 + T]$.
So we have the local uniqueness.

Gronwall's Lemma could be used to show the continuous dependence from initial data. Now assume that we have U_b, b, B, L_A, T like before, $S(t)$ is a solution of (11) on $[0, T]$, $S(0) = S_0$. Continuous dependence on initial data for $S(t)$ means that for any $\varepsilon > 0$ it is possible to find $\delta > 0$ that for all $t \in [0, T]$, $\|S(t) - \tilde{S}(t)\| < \varepsilon$ where $\tilde{S}(t)$ is a solution of (11) with initial data $\tilde{S}(0) = \tilde{S}_0$ and $\|S_0 - \tilde{S}_0\|_1 < \delta$.
◦ We know that $\|S(t) - S_0\|_1 \leq BT < b$. If $\delta < b - BT$ then it is possible to show that $\tilde{S}(t)$ stays inside U_b.
◦ Like before, in **U1**, $\|\tilde{S}(t_*) - S_0\|_1 = b$, $\|\tilde{S}(t) - S_0\| < b$ for $t \in [0, t_*)$

$$\|\tilde{S}(t_*) - S_0\| \leq \|\tilde{S}_0 - S_0\| + \int_{0}^{t_*} \|A(\tilde{S}(\tau))\|_1\, d\tau < b - BT + BT = b$$

and contradiction implies that $\tilde{S}(t)$ stays inside U_b for $t \in [0, T]$).•
For given $\varepsilon > 0$ let $\delta < \min\{b - BT, \varepsilon^* / \exp(L_A T)\}$ then

$$S(t) - \tilde{S}(t) = S_0 - \tilde{S}_0 + \int_{0}^{t} [A(S(\tau)) - A(\tilde{S}(\tau))]\, d\tau$$

$$\|S(t) - \tilde{S}(t)\|_1 \leq \|S_0 - \tilde{S}_0\|_1 + \int_{0}^{t} \|A(S(\tau)) - A(\tilde{S}(\tau))\|_1\, d\tau$$

$$\leq \|S_0 - \tilde{S}_0\|_1 + L_A \int_0^t \|S(\tau) - \tilde{S}(\tau)\|_1 \, d\tau$$

Using Gronwall's lemma with $\delta_1 = \|S_0 - \tilde{S}_0\|_1$, $\delta_2 = L_A$, $\phi(t) = \|S(t) - \tilde{S}(t)\|_1$, $a = T$, $\psi(t) = 1$ we get:

$$\|S(t) - \tilde{S}(t)\|_1 \leq \|S_0 - \tilde{S}_0\|_1 e^{L_A t} < \varepsilon \qquad \forall t \in [0, T] \quad \bullet$$

And this means continuous dependence on initial data for solution $S(t)$ on $[0, T]$.

2.3.3 Expansion of the solution until it reaches zero.

Function S_0 is bounded away from zero if there is a positive constant S_* that $S_0(x) \geq S_*$ almost everywhere; or in other words there is a constant $S_* > 0$ such that $\mu\{x \in \Omega : S_0(x) < S_*\} = 0$. Opposite: function reaches zero if for any positive constant S_*, $\mu\{x \in \Omega : S_0(x) < S_*\} > 0$.

In ODE a solution $y(t)$ existing on $[0, T]$ can be continued until it leaves domain D with "regular" properties of $f(t, y)$ (in D $f(t, y)$ is continuous, Lipschitz continuous on y). The "last" point $(T, y(T))$ being inside D can be used as a "new" initial point, and this procedure can be applied several times. In our case it is also possible to continue from the point $(T, S(T))$, since $S(T)$ staying inside U_b (from **U1**, $S(T)$ is bounded away from zero and we can find "new" U_b for $S_0^{(1)} = S(T)$.) Let use a new notation with $T_0 = 0$, $S_0^{(0)} = S_0$, T is substituted by ΔT_1, $T_1 = T_0 + \Delta T_1$, $S_0^{(1)} = S(T_1)$ and so on ...:

$$\begin{array}{ccccccccccc} T_0 = 0 & & T_1 & & T_2 & & & & T_n & & T_{n+1} \\ S_0^{(0)} = S_0 & \Delta T_1 & S_0^{(1)} & \Delta T_2 & S_0^{(2)} & \cdots & & \Delta T_n & S_0^{(n)} & & S_0^{(n+1)} \end{array} \quad \cdots \qquad (20)$$

In the previous discussion, the most important point was that there exists some $T > 0$ from (19), provided $S_0 > 0$ but the choice of U_b and T was not fixed. Here we need to fix them to exclude, for example, mean-less choice of ΔT_n in (20) with $\Delta T_n = \min\{b_n / B_n, 1/L_A^{(n)}\}/n^2$.

Remark The optimal choice is not our purpose.

Let us use the following notation:

$$d[S] = \operatorname*{ess\,inf}_{x \in \Omega} S(x), \quad \mu(d) = \inf_{s \in [d,1]} K(s), \quad \mu[S] = \mu(d[S]), \quad d_n = d[S_0^{(n)}] \qquad (21)$$

We know that $\Delta T_n > 0$ for all $n \in \mathbf{N}$ (because of **U1** and $S(T_n) \geq d_n > 0$), but ΔT_n may become smaller and smaller when $n \to \infty$. If $T^* = \sup_n T_n = \lim_{n \to \infty} T_n < \infty$ then for $t \geq T^*$ we cannot define a solution $S(t)$ by the sequence (20). Our purpose is to estimate ΔT_n from below in order to clarify the situation with T^*.

Determination of one possible process (20).

$n \in \mathbf{N}$. $S(T_n) = S_0^{(n)} \geq d_n > 0$. Let $b_n = d_n / 2C^B$ then $\forall S \in U_{b_n}$, a.e:

$$|S(x) - S_0^{(n)}(x)| \leq C^B \|S - S_0^{(n)}\|_1 \leq C^B b_n = \frac{d_n}{2},$$

$$|S(x)| \geq |S_0^{(n)}(x)| - |S_0^{(n)}(x) - S(x)| \geq d_n - \frac{d_n}{2} = \frac{d_n}{2} > 0.$$

$\Rightarrow \quad d[S] \geq \frac{d_n}{2} \ \forall S \in U_{b_n}$.

On the other hand we do not want to have b_n too large, so let

$$b_n = \frac{1}{2C^B} \min\{d_n, 1\} \qquad (22)$$

U_{b_n} is defined. For given U_{b_n} we can choose

$$\Delta T_n = 0.9 \min\{1/L_A^{(n)}, b_n/B_n\} < \min\{1/L_A^{(n)}, b_n/B_n\}. \tag{23}$$

To estimate ΔT_n from below we need to estimate the positive constants $L_A^{(n)}$, B_n, b_n; two first from above and the last from below.
(21) implies that $\mu(d)$ is monotone: $\varepsilon \le d_n \Rightarrow \mu(\varepsilon) \le \mu(d_n)$. Then $\forall S \in U_{b_n}$, $\mu[S] = \mu(d[S]) \ge \mu(d_n/2) \Rightarrow$ estimation of the ellipticity constant for U_{b_n} from below:

$$E_n = \inf_{S \in U_{b_n}} E[S] = \inf_{S \in U_{b_n}} \min\{\mu[S]L, \phi_0\} \ge \min\left\{\mu\left(\frac{d_n}{2}\right)L, \phi_0\right\}.$$

From **A1**:

$$B_n = \Big(\|KP'\|_\infty(\|S_0^{(n)}\|_1 + b_n) + \rho g \|K\|_\infty \mu(\Omega)\Big) \Big/ E_n. \tag{24}$$

Remark We need to take into account a possibility: $\|S_0^{(n)}\|_1 \to \infty$ when $n \to \infty$, but we can control this by the choice of b_n from (22):

$$\|S_0^{(n)}\|_1 + b_n \le \|S_0^{(n-1)}\|_1 + b_{n-1} + b_n \le \ldots \le \|S_0^{(0)}\|_1 + \sum_{k=0}^{n} b_k \le \|S_0^{(0)}\|_1 + \frac{n+1}{2C^B}.$$

Dependence B_n on n has the form $B_n \le (\beta^1 n + \beta^2)/E_n$, where positive constants β^1, β^2 can be expressed from (24).

$$L_A^{(n)} = \big(L_{KP'}C^B(\|S^0\|_1 + b) + \|KP'\|_\infty + \rho g L_K C^B \mu(\Omega) + L_K C^B L B_n\big) \big/ E_n. \tag{25}$$

In simplified form (with positive constants α^1, α^2, α^3, β^1, β^2 which do not depend on n):

$$L_A^{(n)} \le \big(\alpha^1 n + \alpha^2 + \alpha^3(\beta^1 n + \beta^2)/E_n\big) \big/ E_n.$$

For the sequence of strictly positive numbers d_n we can distinguish two possibilities:

$$1)\ d_n \ge \varepsilon > 0 \text{ for all } n \qquad \text{or} \qquad 2)\ \liminf_{n\to\infty} d_n = 0.$$

1) In the first case

$$E_n \ge \min\{\mu(\varepsilon/2)L, \phi_0\} = E_\varepsilon > 0,$$

$$B_n \le (\beta^1 n + \beta^2)/E_\varepsilon, \qquad L_A^{(n)} \le (\gamma^1 n + \gamma^2)/E_\varepsilon, \qquad b_n \ge \varepsilon/2C^B.$$

$$\Delta T_n = 0.9 \min\{1/L_A^{(n)}, b_n/B_n\} \ge 0.9 \min\left\{\frac{E_\varepsilon}{\gamma^1 n + \gamma^2}, \frac{\varepsilon}{2C^B(\beta^1 n + \beta^2)}\right\}$$

Beginning from some number n_0 one from $\{E_\varepsilon/(\gamma^1 n + \gamma^2), \varepsilon/2C^B(\beta^1 n + \beta^2)\}$ is always smaller than another,

$$T^* = \sum_{n=0}^{\infty} \Delta T_n \ge \sum_{n=n_0}^{\infty} \Delta T_n \ge \min\left\{\sum_{n=n_0}^{\infty} \frac{0.9\,E_\varepsilon}{\gamma^1 n + \gamma^2}, \sum_{n=n_0}^{\infty} \frac{0.9\,\varepsilon}{2C^B(\beta^1 n + \beta^2)}\right\} = \infty.$$

both rows do not converge. In this case $T^* = \infty$ and we can continue the solution til any positive value t.
2) In the second case the solution "reaches zero". When $d(S) = 0$, the equation (6) loses ellipticity, and for such S, operator $A(S)$ may be undefined. We can determine

the solution $S(t)$ on $[0, T^*)$ by sequence (20), but we do not know if T^* is finite or not (it may also be infinite like in the first case).

In both cases the solution $S(t)$ can be determined on $[0, T^*)$ by infinite process (20). In each segment $[T_n, T_{n+1}]$, the solution $S(t)$ is unique $\Rightarrow$ $S(t)$ is unique on $[0, T^*)$. For any given $t \in [0, T^*)$, the solution $S(t)$ depends continuously on initial value S_0 (there exists finite number n that $T_n > t$ and we can apply continuous dependence on each segment $[T_{k-1}, T_k]$, beginning from the last $k = n \dots 1$.)
In the case 2) we have analogy with ODE case when a solution leaves domain of "good" properties of $f(t, y)$.
Remark We need to mention that S being a saturation cannot be greater than 1 – another critical boundary with which we may deal in a same way like with 0: $T^* = \infty$ or S "reaches" the critical boundaries (0 or 1).

2.3.4 Solution as a function of two variables x and t.

Until now we were dealing with the abstract function $S \in C([0, T] \to H^1(\Omega))$. In this section we are going to find some properties of the solution as a "normal" function of two variables to obtain in some way a connection between (11) and (6).

Let $y \in C([0, T] \to H^1(\Omega))$ and I be an embedding operator from $H^1(\Omega)$ to $C_B(\Omega)$. $Iy \in C([0, T] \to C_B(\Omega))$ We will also consider that I: $H^1(\Omega) \to C_B(\Omega) \bigcap H^1(\Omega)$, in other words $Iy(t) \in H^1$, $Iy(t) = y(t)$ in H^1 and for example $A(Iy(t)) = A(y(t))$.

So $\forall t$, $y(t)$ being a function from $H^1(\Omega)$ has a continuous representative $y(t, \cdot) := Iy(t) \in C_B \bigcap H^1$. $y(t, x)$ is a real function from t and x. In every point $x \in \Omega$ it is uniquely defined.

In the section "Another formulation ..." the space $C_B(\Omega)$ was already mentioned. $C_B(\Omega)$ is a Banach space of bounded continuous functions (not necessarily uniformly continuous) under the norm

$$\|y\|_{C_B} = \sup_{x \in \Omega} |y(x)|$$

Remark In one dimensional case more regular $C(\bar{\Omega})$ can be used instead of $C_B(\Omega)$.
C1. $y(t, x)$ is continuous on $[0, T] \times \Omega$.
$\circ$ $(t, x) \in [0, T] \times \Omega$. $B_r(x) \subset \Omega$ - a ball with center x and some positive radius r.
For any $\varepsilon > 0$ exists $\delta < r$ that:
a) $\|y(t_1, \cdot) - y(t, \cdot)\|_1 < \varepsilon / 2C_B$ when $|t_1 - t| < \delta$

$$\Rightarrow \qquad \sup_{x \in \Omega} |y(t_1, x) - y(t, x)| \le C_B \|y(t_1, \cdot) - y(t, \cdot)\|_1 < \varepsilon / 2.$$

b) $y(t, x)$ is continuous in x: $|y(t, x_1) - y(t, x)| < \varepsilon / 2$ when $|x_1 - x| < \delta$.
For all (t_1, x_1): $|x_1 - x| < \delta$, $|t_1 - t| < \delta$:

$$|y(t_1, x_1) - y(t, x)| \le |y(t_1, x_1) - y(t, x_1)| + |y(t, x_1) - y(t, x)| < \varepsilon/2 + \varepsilon/2 = \varepsilon \bullet$$

C2. $y(\cdot) \in C\left([0, T] \to H^1(\Omega)\right)$ Then $Iy(\cdot) \in C\left([0, T] \to C_B(\Omega)\right)$ and

$$I \int_0^t y(\tau)\, d\tau = * \int_0^t Iy(\tau)\, d\tau$$

Remark The Integral in the right hand side is in the sense of the Banach space $E = C_B(\Omega)$ for which properties **Y1**–**Y7** are also valid. Integrals in C_B we will mark by $*$ before the integral.
$\circ$ a) I-continuous, $y(\cdot) \in C\left([0, T] \to H^1(\Omega)\right) \Rightarrow Iy(\cdot) \in C\left([0, T] \to C_B(\Omega)\right)$. Then

exists $*\int_0^t Iy(\tau)\,d\tau$.
b)I – linear $\Rightarrow$

$$I(\sum_i y(\tau_i)\Delta\tau_i) = \sum_i Iy(\tau_i)\Delta\tau_i.$$

Right side tends to $\int_0^t Iy(\tau)\,d\tau$. Since I is continuous the left side tends to $I\int_0^t y(\tau)\,d\tau$ •

Consider $S \in C\left([0,T] \to H^1\right)$ a solution of (11). It satisfies (18).
From **C2**: $IS(\cdot) \in C\left([0,T] \to C_B\right)$,

$$IS(t) = IS_0 + *\int_0^t IA(S(\tau))\,d\tau.$$

Using $IA(S(\cdot)) \in C\left([0,T] \to C_B\right)$ and **Y6** we can conclude that $\forall t \in [0,T]$, $IS(t)$ is differentiable in the sense of C_B and **Def 2** (is marked by *):

$$*\frac{d}{dt}IS(t) = IA(S(t)) = IA(IS(t)).$$

A function of two variables $S(t,\cdot) := IS(t)$ is continuous on (t,x) from **C1**. Existence of $*\frac{d}{dt}IS(t)$ implies existence of classical partial derivative from $S(t,x)$ on t: $\frac{\partial}{\partial t}S(t,\cdot) = *\frac{d}{dt}IS(t)$. Hence,

$$\frac{\partial}{\partial t}S(t,x) = IA(S(t,\cdot))(x). \tag{26}$$

Moreover, $\frac{\partial}{\partial t}S(t,x)$ is continuous since the right hand side is continuous on $[0,T]\times\Omega$.
We can summarize properties of $S(t,x)$:
C3. $S(t,x)$ - is continuous on $[0,T]\times\Omega$ and $S(t,\cdot)$ is continuous on $[0,T]$ in $H^1(\Omega)$ $(S(\cdot,\cdot) \in C\left([0,T] \to H^1(\Omega)\right))$.
C4. Exists (continuous) $\frac{\partial}{\partial t}S(t,x)$ on $[0,T]\times\Omega$ and $\frac{\partial}{\partial t}S(t,\cdot) \in H^1(\Omega)$.
C5. $S(t,x)$ satisfies (26), $S(0,x) = IS_0(x) \geq d[S_0] > 0$ and it looks possible to say that $S(t,x)$ is a weak solution on x of (6) (in the variational sense).

Assume that function $\tilde{S}(t,x)$ satisfies **C3,C4,C5**. We are going to show that $\tilde{S}(t,x)$ has to be equal to $S(t,x)$ defined before.
Remark Existence of $\frac{\partial}{\partial t}\tilde{S}$ may be insufficient for existence of $*\frac{d}{dt}\tilde{S}$, $\frac{d}{dt}\tilde{S}$.
◦ From **C3** we know that $\tilde{S}(t,\cdot)$ is continuous in H^1. So $A(\tilde{S}(t,\cdot))$, $IA(\tilde{S}(t,\cdot))$ are also continuous in H^1 and C_B (while $d[\tilde{S}(t,\cdot)] > 0$). Continuity of $\frac{\partial}{\partial t}\tilde{S}(t,x)$ can be obtained as consequence of **C5**, previous sentence and **C1** (it is not necessary to have it in **C2** as a condition). $D(t) = \int_0^t A(\tilde{S}(\tau,\cdot))\,d\tau$ is well defined in H^1.

$$ID(t) = *\int_0^t IA(\tilde{S}(\tau,\cdot))\,d\tau \overset{\mathbf{C5}}{=} *\int_0^t \frac{\partial}{\partial t}\tilde{S}(\tau,\cdot)\,d\tau \quad \text{is well defined in } C_B$$

We need to show that $ID(t)(x) = \tilde{S}(t,x) - \tilde{S}(0,x)$ (in other words that $*\frac{d}{dt}\tilde{S}$ exists and $\frac{\partial}{\partial t}\tilde{S} = *\frac{d}{dt}\tilde{S}$)
◦ Assume it is not true:

$$ID(t)(x_0) \neq \tilde{S}(t,x_0) - \tilde{S}(0,x_0) = **\int_0^t \frac{\partial}{\partial t}\tilde{S}(\tau,x_0)\,d\tau.$$

Last integral exists in $\mathbf{R}$ since $\frac{\partial}{\partial t}\tilde{S}(t,x)$ is continuous on $[0,T]\times\Omega$. $**-$integral in $\mathbf{R}$.
Let us

$$|ID(t)(x_0) - **\int_0^t \frac{\partial}{\partial t}\tilde{S}(\tau,x_0)\,d\tau| = \varepsilon. \tag{27}$$

Then from integral definitions in $\mathbf{R}$ and C_B we can conclude that there is $\delta > 0$ such that for each partition $\mathcal{T}$ of $[0,t]$ with $\Delta\tau_i < \delta$,

$$\left| \sum_i \frac{\partial}{\partial t} \tilde{S}(\tau_i, x_0) \Delta\tau_i - * * \int_0^t \frac{\partial}{\partial t} \tilde{S}(\tau, x_0)\, d\tau \right| < \frac{\varepsilon}{2},$$

$$\left| \sum_i \frac{\partial}{\partial t} \tilde{S}(\tau_i, x_0) \Delta\tau_i - ID(t)(x_0) \right| \leq \left\| \sum_i \frac{\partial}{\partial t} \tilde{S}(\tau_i, \cdot) \Delta\tau_i - ID(t) \right\|_{C_B} < \frac{\varepsilon}{2}.$$

and we have contradiction with (27) • So

$$* \int_0^t IA(\tilde{S}(\tau, \cdot))\, d\tau = \tilde{S}(t, \cdot) - \tilde{S}(0, \cdot) = \tilde{S}(t, \cdot) - IS_0.$$

From **C3**, there is only one function $\tilde{S}(t) \in H^1$, that $I\tilde{S}(t) = \tilde{S}(t, \cdot)$, and $\tilde{S}(t)$ satisfies (18). Then $\tilde{S}(t)$ is a solution of (11), but it must be unique. So $\tilde{S}(t) = S(t)$, $\Rightarrow$ $\tilde{S}(t, \cdot) = S(t, \cdot)$ •.

Remark C3 can be rather strong condition for continuous function $S(t, x)$.

We can summarize all what we tried to obtain in this section in the following: Suppose the assumptions (9) are satisfied. Then there is $T^* > 0$, finite or infinite such that for any constant $T < T^*$, on $[0, T]$ there exists a unique solution $S(t)$ of (11) in H^1 that continuously depends on initial data S_0. T^* can be estimated from below by ΔT_1 from (23, equality). This solution can be considered as a function of two variables $S(t, x) = IS(t)(x)$ that satisfies **C3, C4, C5** and there is no other function of two variables $\tilde{S}(t, x)$ that satisfies **C3, C4, C5**.

We cannot establish further connection with (4).

3 Numerical methods

In this section our main task is finding the numerical methods for (4). For simplicity we will deal only with homogeneous boundary conditions B.C.1 and B.C.2 for one dimension problem, like in the previous section. Instead of doing discretization of (4) we will do numerics for (11) or the intermediate form (6), assuming that this substitution is reasonable. We are going to use similarity of (11) with ODE case.

The operator $A(S)$ gives a solution u of elliptic problem (6) for known S. Let $A^h(S_h)$ be a numeric approximation of $A(S)$. So $A^h(S_h)$ gives a numeric solution of elliptic problem (6) in Ω for corresponding boundary condition and known S_h. And it is possible to use finite element and finite difference methods to approximate $A(S)$.

In finite element approach, $S_h \in H_h^1$, $A^h : H_h^1 \to V_h$, where V_h and H_h^1 are finite dimension subspaces of Hilbert spaces V and H^1. (We remind that $V = H^1(\Omega)$ for B.C.1 and $V = H_0^1(\Omega)$ for B.C.2, $\Omega = (0, l)$).

Using finite difference schemes we have some grid G in Ω,
$G = \{x_i \in \Omega \quad i = 1, \dots m\}$. Then $S_h \in \mathbf{R}^m$ and $A^h : \mathbf{R}^m \to \mathbf{R}^m$.

Assume that we have some numerical approximation $A^h : W_h \to U_h \subseteq W_h$, no matter what W_h, U_h are, what boundary condition we have and which numeric method we use for A^h.

Consider following analog of (11):

$$\frac{d}{dt} S_h(t) = A^h(S_h(t)), \qquad S_h(t) \in W_h, \qquad S_h(0) = S_h^0. \tag{28}$$

This equation is similar to ODE $\frac{d}{dt} y(t) = f(t, y(t))$, $y(0) = y_0$ or to System of ODE $\frac{d}{dt} \bar{x}(t) = \bar{f}(t, \bar{x}(t))$, $\bar{x}(0) = \bar{x}_0$.

Remark We can introduce a dependence of A^h on t. In the case of homogeneous boundary conditions and coefficients in (6) that depend only on S and x, it is only a formalism and $A^h(S_h, t) = A^h(S_h)$. But in more general problems A^h depends on t.

There are many numerical methods known for ODE, and many of them can be employed for (28).

Let us consider an equidistant grid in t with some step Δt and following notations: $t^0 = 0$, $t^{j+1} = t^j + \Delta t$, $S_h^j = S_h(t^j)$, $j = 0 \dots N$, $\Delta t = T/N$.

Example 1 Runge-Kutta methods.

$$S_h^{j+1} = S_h^j + \Delta t \sum_{k=1}^{p} b_k A_k$$

$$A_1 = A^h(S_h^j, t^j),\ A_2 = A^h(S_h^j + \Delta t a_{21} A_1, t^j + c_2 \Delta t), \dots$$

$$A_p = A^h\left(S_h^j + \Delta t \sum_{k=1}^{p-1} a_{pk} A_k, t^j + c_p \Delta t\right).$$

Constants a_{lk}, c_k, b_k determine the Runge-Kutta method.

Euler method $p = 1$

$$S_h^{j+1} = S_h^j + \Delta t A^h(S_h^j, t^j). \tag{29}$$

Improved Euler method $p = 2$

$$S_h^{j+1} = S_h^j + \Delta t A^h\left(S_h^j + \frac{\Delta t}{2} A^h(S_h^j, t^j), t^j + \frac{\Delta t}{2}\right). \tag{30}$$

Euler-Cauchy method $p = 2$

$$S_h^{j+1} = S_h^j + \Delta t \left[A^h(S_h^j, t^j) + A^h \left(S_h^j + \Delta t A^h(S_h^j, t^j), t^j + \Delta t \right) \right] /2. \qquad (31)$$

Fourth order Runge-Kutta method $p = 4$

$$S_h^{j+1} = S_h^j + \Delta t \left(\frac{1}{6} A_1 + \frac{1}{3} A_2 + \frac{1}{3} A_3 + \frac{1}{6} A_4 \right) \qquad (32)$$

$$A_1 = A^h(S_h^j, t^j), \qquad A_2 = A^h \left(S_h^j + \tfrac{1}{2}\Delta t A_1, t^j + \tfrac{1}{2}\Delta t \right),$$
$$A_3 = A^h \left(S_h^j + \tfrac{1}{2}\Delta t A_2, t^j + \tfrac{1}{2}\Delta t \right), \quad A_4 = A^h \left(S_h^j + \Delta t A_3, t^j + \Delta t \right).$$

Example 2 The multi-step method.

$$S_h^{j+p} = S_h^{j+q} + \Delta t \sum_{k=0}^{p} b_k A^h(S_h^{j+k}, t^{j+k})$$

where p, q, b_k $k = 0, \ldots, p$ are particular method's parameters.

Remark It is also possible to use implicit schemes for ODE.

One approach for constructing a numerical method for (4) is to combine some numeric method for Elliptic Boundary Value problem with some numerical method for Ordinary Differential Equation. Both choices can be rather independent from each other.

Next we will discuss a combination of Euler method in "t - direction" with abstract finite element method in "x - direction"; describe one implementation of finite element method for $A^h(S)$ where V_h is a space of piece-wise linear functions; describe another approximation $A^h(S)$ from finite differences approach. At the end we present some computational experiment results for different constants $L > 0$, two types of homogeneous boundary conditions B.C.1, B.C.2; comparison with results for Richard's equation ($L = 0$) and comparison the results for the same problem but obtained on nested sequence of grids.

3.1 Finite elements – Euler method

For the most simplest method we will try to get convergence. The operator A was defined in H^1 by variational approach. So it is natural to use finite element method to get it's approximation $A^h(S)$.

Assume that we have a differential problem (11) with S_0 bounded away from zero. We can choose U_b with constants b, B, L_A, C, E like before. Let $S(t)$ be a solution of (11) on $[0, T]$, where T is from (19) for chosen U_b. $S(t)$ lies inside U_b not near than $b - BT$ from the boundary ∂U_b.

Assume that we have a sequence of finite dimensional subspaces $\{V_h\}$, where every next element contains all previous; parameter h is one from a monotone decreasing sequence $\{h_k\}$, $\lim\limits_{k\to\infty} h_k = 0$. When $h_k \to 0$, the dimension of V_{h_k} increases and they exhaust all V ($\forall \varepsilon > 0, \forall u \in V \ \exists V_h$ from the sequence that: $\inf\limits_{v_h \in V_h} \|u - v_h\|_1 < \varepsilon$).

To get approximation $u_h = A^h(S_h)$ of $u = A(S_h)$ by finite element method for chosen V_h, we need to find $u_h \in V_h$ that (see (8)):

$$a^*[S_h](u_h, v_h) = a[S_h](u_h, v_h) + (u_h, v_h)_{0,\phi} = l[S_h](v_h) \qquad \forall v_h \in V_h$$

$\dim V_h = d$. $v_1 \dots v_d$ – basis in V_h.
To find $u_h = \sum_{i=1}^{d} u_h^i v_i$, we have to solve a linear algebraic system of equations:

$$\sum_{i=1}^{d} a^*[S_h](v_i, v_j) u_h^i = l[S_h](v_i), \qquad j = 1 \dots d, \tag{33}$$

with positive definite matrix $\{a^*[S_h](v_i, v_j)\}_{ij}$ provided $a^*[S_h]$ is V-elliptic. So the approximate solution u_h exists for all $S_h \in U_b \subset H^1$.

The difference between approximate and exact solutions can be estimated with the help of **Cea lemma** (see [5] p. 54 or [3] part b, p.118):

$$\|u - u_h\|_V \le \frac{C}{E} \inf_{v_h \in V_h} \|u - v_h\|_V, \qquad u = A(S_h), \quad u_h = A^h(S_h), \quad S_h \in U_b \cap V_h.$$

Remark It is difficult to expect that we can find V_h uniformly closed to the set $A(U_b) = \{u = A(S_h) : S_h \in U_b\}$ when $h \to 0$, in other words that for any $\varepsilon > 0$ exists V_h that $\forall u \in A(U_b)$, $\inf\limits_{v \in V_h} \|u - v\|_V < \varepsilon$.

Some notations:

$$\begin{array}{lcll} S_h^j & = & S_h(t^j) & - \text{ approximate solution at time } t^j. \\ S^j & = & S(t^j) & - \text{ exact solution at time } t^j. \\ \Delta^j & = & S^j - S_h^j & - \text{ error of approximation.} \end{array}$$

$S^j, S_h^j, \Delta^j \in H^1(\Omega)$.

$$S^{j+1} = S^j + \int_{t^j}^{t^{j+1}} A(S(\tau))\, d\tau \approx S^j + A(S^j)\Delta t$$

$$\|S^{j+1} - S^j - A(S^j)\Delta t\|_1 = \| \int_{t^j}^{t^{j+1}} \left[A(S(\tau)) - A(S^j)\right]\, d\tau\|_1 \le$$

$$\le \int_{t^j}^{t^{j+1}} \|A(S(\tau)) - A(S^j)\|_1\, d\tau \le L_A B \frac{(t^{j+1} - t^j)^2}{2} \tag{34}$$

we used $\quad \|A(S(\tau)) - A(S^j)\|_1 \le L_A \|S(\tau) - S^j\|_1 \le L_A B|\tau - t^j|$.

The Euler method: $S_h^{j+1} = S_h^j + \Delta t A^h(S_h^j)$.
Remark $S_h^j \in V_h + S_h^0$ for all j.

$A^h(S_h^j) = A(S_h^j) + e_h^j$, e_h^j is an error of finite element method, and can be estimated by **Cea lemma**. Due to the difficulties noticed in the last remark, we would like to find V_h uniformly closed to $A(S(t))$, where $S(t)$ is exact solution.

For some small $\varepsilon_1 > 0$, let divide $[0, T]$ into l parts by points $\tau_i = i\Delta\tau$, $i = 0 \dots l$, and $\Delta\tau \le \varepsilon_1 / 2L_A B$. Then

$$\|A(S(t)) - A(S(\tau_i))\|_1 \le L_A \|S(t) - S(\tau_i)\|_1 \le L_A B|t - \tau_i| < L_A B \Delta\tau \le \frac{\varepsilon_1}{2} \tag{35}$$

where $t \in [0, T]$, τ_i is the nearest point to t ($\Delta\tau$ is not the same with Δt in Euler method). l is a finite number $\Rightarrow$ we can choose V_h such that

$$\max_{i=0 \dots l} \inf_{v_h \in V_h} \|A(S(\tau_i)) - v_h\|_1 < \frac{\varepsilon_1}{2}. \tag{36}$$

(36) are also true for all elements from $\{V_h\}$ that follow the chosen subspace (because $V_{h_k} \subset V_{h_{k+1}}$).

Together (35), (36) give $\inf\limits_{v_h \in V_h} \|A(S(t)) - v_h\|_1 < \varepsilon_1$ for all $t \in [0, T]$.

We can estimate e_h^j:

$$\|e_h^j\|_1 = \|A(S_h^j) - A^h(S_h^j)\|_1 \leq \frac{C}{E} \inf_{v_h \in V_h} \|A(S_h^j) - v_h\|_1 \leq$$

exists $v_h^* \in V_h$ that $\|A(S^j) - v_h^*\|_1 < \varepsilon_1 \Rightarrow$

$$\leq \frac{C}{E} \left(\|A(S_h^j) - A(S^j)\|_1 + \|A(S^j) - v_h^*\|_1 \right) \leq \frac{C}{E}(L_A \|\Delta^j\|_1 + \varepsilon_1). \qquad (37)$$

Now we can estimate the error Δ^{j+1} from Δ^j:

$$\Delta^{j+1} = S^{j+1} - S_h^{j+1} = S^{j+1} - S_h^j - A(S_h^j)\Delta t - e_h^j \Delta t =$$

$$= S^{j+1} - S^j - A(S^j)\Delta t + S^j + A(S^j)\Delta t - S_h^j - A(S_h^j)\Delta t - e_h^j \Delta t.$$

$\|\Delta^{j+1}\|_1 \leq \|S^{j+1} - S^j - A(S^j)\Delta t\|_1 + \|S^j - S_h^j\|_1 + \Delta t \|A(S^j) - A(S_h^j)\|_1 + \Delta t \|e_h^j\|_1 \leq$

From (34) we estimate the first term, from (37) the last.

$$\leq L_A B \frac{\Delta t^2}{2} + \|\Delta^j\|_1 + \Delta t L_A \|\Delta^j\|_1 + \Delta t \frac{C}{E}(L_A \|\Delta^j\|_1 + \varepsilon_1) =$$

$$= \|\Delta^j\|_1 \left(1 + \Delta t L_A (1 + \frac{C}{E})\right) + \Delta t \left(L_A B \frac{\Delta t}{2} + \frac{C}{E} \varepsilon_1 \right) = \alpha \|\Delta^j\|_1 + \Delta t D.$$

where $\alpha = 1 + \Delta t L_A (1 + \frac{C}{E}) > 1, \qquad D = L_A B \frac{\Delta t}{2} + \frac{C}{E} \varepsilon_1.$

$$x_j = \|\Delta^j\|_1, \qquad x_{j+1} \leq \alpha x_j + \Delta t D.$$

The initial error $x_0 = \|S^0 - S_h^0\|_1$ is zero if we know the exact value of $S^0 \in H^1$ (then we can $S_h^0 := S^0$, in this case $S_h^j \in V_h + S^0$). It is not necessary to take approximation S_h^0 from some H_h^1.

Now consider that exact initial function $S^0 \in H^1$ is unknown, the measured value $\tilde{S}^0 \in H^1$ has a measurement error $e_M = \|S^0 - \tilde{S}^0\|_1$. Additionally, if it is convenient to use functions from H_h^1 instead of H^1, then we have an approximation error $e_A = \|\tilde{S}^0 - S_h^0\|_1$. (It can be convenient since integrals in $a^*[S_h]$ may be too complicated when $S_h \in V_h + S_h^0$, for arbitrary $S_h^0 \in H^1$).

Remark It is not necessary to have either $V_h = H_h^1$ for B.C.1 ($V = H^1$) or $V_h \subset H_h^1$ for B.C.2 ($V = H_0^1$); it can be so only if convenient.

$$x_0 = \|S^0 - S_h^0\|_1 \leq e_A + e_M = e$$

$x_{j+1} \leq \alpha x_j + \Delta t D$, everything is positive. Let $y_0 = x_0$, $y_{j+1} = \alpha y_j + \Delta t D$.
$\{y_j\}$ is an upper boundary for x_j, for all j: $y_j \geq x_j$.
$y_1 = \alpha y_0 + \Delta t D$, $y_2 = \alpha^2 y_0 + \alpha \Delta t D + \Delta t D$. For arbitrary $j \in \{0, \ldots N\}$:

$$y_j = \alpha^j y_0 + \alpha^{j-1} \Delta t D + \ldots + \Delta t D = \alpha^j y_0 + \Delta t D \sum_{i=0}^{j-1} \alpha^i = \alpha^j y_0 + \Delta t D \frac{\alpha^j - 1}{\alpha - 1}.$$

$$\alpha^N = \left(1 + \Delta t L_A \left(1 + \frac{C}{E}\right)\right)^N = \left(1 + \frac{L_A T(1 + C/E)}{N}\right)^N \le e^{L_A T(1+C/E)} = C_1.$$

$\frac{\Delta t}{\alpha - 1} = [L_A(1 + C/E)]^{-1} = C_2$

$$x_j \le y_j \le y_N \le C_1 x_0 + C_2(C_1 - 1)D = O(\varepsilon_1 + e + \Delta t).$$

And for any $\varepsilon > 0$ there exist so small ε_1, e, Δt that $\|\Delta^j\|_1 < \varepsilon$ for all $j = 0 \dots N$. ε_1 can be made arbitrary small for sufficiently large V_h from $\{V_n\}$. And we assume that measurement error e_M and approximation error e_A can be made arbitrary small or even zero.

The last thing: if the given ε is larger than $b - BT$, then $\varepsilon := b - BT$. We need this to guarantee that S_h^j stays in U_b to use properties of $A(S)$.

Remark Convergence in h is without order, in Δt – first order.

Remark It seems not reasonable to try to get higher order in Δt for other Runge-Kutta method using the same means. We have only Lipschitz continuity on $A(S(t))$, we do not know if $\frac{d}{dt}A(S(t))$ exists and we cannot get better integral approximation in (34) than the second order. For instance $|x|$ is a Lipschitz continuous, but not differentiable function on $[-h, h]$ and integral approximations give only $O(h^2)$ error, not $O(h^3)$.

3.2 One possible $A^h(S)$ by finite element method.

In the previous section spaces $\{V_h\}$ were not specified. Here we choose V_h as a space of continuous functions, linear between grid points x_i, $x_i = ih$, $i = 0 \dots n$, $h = l/n$.

The standard basis in H_h^1 is:

$$\psi_i = \begin{cases} \frac{1}{h}(x - x_{i-1}) & \text{if} \quad x \in [x_{i-1}, x_i] \\ -\frac{1}{h}(x - x_{i+1}) & \text{if} \quad x \in [x_i, x_{i+1}] \\ 0 & \text{otherwise} \end{cases} \qquad i = 1 \dots n-1.$$

And additionally two functions:

$$\psi_0 = \begin{cases} -\frac{1}{h}(x - x_1) & , x \in [x_0, x_1] \\ 0 & \text{otherwise} \end{cases}, \quad \psi_n = \begin{cases} \frac{1}{h}(x - x_{n-1}) & , x \in [x_{n-1}, x_n] \\ 0 & \text{otherwise} \end{cases}$$

When we have B.C.1 then we use $V_h = H_h^1$ with basis $v_i = \psi_{i-1}$, $i = 1 \dots d$, $d = n+1$. For B.C.2 $V_h \subset H_h^1$ has basis $v_i = \psi_i$, $i = 1 \dots d$, $d = n-1$. In general $A_h : H^1 \to V_h$, but we will use only $A_h : H_h^1 \to V_h$: if the initial function $S_h^0 \in H_h^1$ hence all other approximate solutions S_h^j be from $S_h^0 + V_h \subset H_h^1$.

To find a value $u_h = A^h(S_h)$ we need to solve a System of Linear Algebraic Equations (33) with positive definite matrix $\{a_{ij}\}$ and right hand side vector b, $a_{ij} = a^*[S_h](v_i, v_j)$, $b_i = l[S_h](v_i)$, $i, j \in \{1 \dots d\}$. Our purpose here is to simplify expressions for a_{ij} and b_i using formulas for v_i and piece-wise linearity of $S_h \in H_h^1$. Then we can denote $S_i = S_h(x_i)$. $S_h(x)$ is uniquely defined by these numbers. Also $u_i = u_h(x_i)$.

To cover both boundary conditions we will calculate $a_{ij} = a^*[S_h](\psi_i, \psi_j)$, $b_i = l[S_h](\psi_i)$, $i, j = \{0 \dots n\}$. This matrix coincides with those for B.C.1; the matrix for B.C.2 could be obtained by deleting rows and columns with number 0 and n.

First we notice that matrix $\{a_{ij}\}$ is tridiagonal, symmetric ($|i-j| > 1 \Rightarrow \psi_i \psi_j \equiv 0$, $\frac{d\psi_i}{dx}\frac{d\psi_j}{dx} \equiv 0$.)

$$\frac{d\psi_i}{dx}\frac{d\psi_i}{dx} = \begin{cases} \frac{1}{h^2} & \text{on} \quad [x_{i-1}, x_{i+1}] \\ 0 & \text{otherwise} \end{cases} \qquad i = 1 \dots n-1.$$

for $i = 0$ and $i = n$ this expression is true while $x \in (0, l)$.

$$\frac{d\psi_i}{dx}\frac{d\psi_{i-1}}{dx} = \begin{cases} -\frac{1}{h^2} & \text{on} \quad [x_{i-1}, x_i] \\ 0 & \text{otherwise} \end{cases} \qquad i = 1 \dots n.$$

$$\frac{d\psi_i}{dx}\frac{d\psi_{i+1}}{dx} = \begin{cases} -\frac{1}{h^2} & \text{on} \quad [x_i, x_{i+1}] \\ 0 & \text{otherwise} \end{cases} \qquad i = 0 \dots n-1.$$

$$\psi_i\psi_i = \begin{cases} \frac{1}{h^2}(x - x_{i-1})^2 & x \in [x_{i-1}, x_i] \\ \frac{1}{h^2}(x - x_{i+1})^2 & x \in [x_i, x_{i+1}] \\ 0 & \text{otherwise} \end{cases} \qquad i = 1 \dots n-1.$$

for $i = 0$ and $i = n$ the last expression is true while $x \in (0, l)$.

$$\psi_i\psi_{i-1} = \begin{cases} -\frac{1}{h^2}(x - x_{i-1})^2 + \frac{1}{h}(x - x_{i-1}) & x \in [x_{i-1}, x_i] \\ 0 & \text{otherwise} \end{cases} \qquad i = 1 \dots n.$$

$$\psi_i\psi_{i+1} = \begin{cases} -\frac{1}{h^2}(x - x_{i+1})^2 - \frac{1}{h}(x - x_{i+1}) & x \in [x_i, x_{i+1}] \\ 0 & \text{otherwise} \end{cases} \qquad i = 0 \dots n-1.$$

We remind the integral expressions for $a^*[S_h]$, $l[S_h]$:

$$a_{ij} = a^*[S_h](\psi_i, \psi_j) = \int_0^l K(S_h(x)) L \frac{d\psi_i}{dx}\frac{d\psi_j}{dx}\,dx + \int_0^l \phi\psi_i\psi_j\,dx$$

$$b_i = l[S_h](\psi_i) = \int_0^l F(S_h(x))\frac{d\psi_i}{dx}\,dx = \int_0^l K(S_h(x))\left(\frac{\partial P(S_h(x))}{\partial x} - \rho g\right)\frac{d\psi_i}{dx}\,dx$$

Integrands are not zero only on a small interval with the length h or $2h$. For a_{ii} for $i = 1 \dots n-1$:

$$a_{ii} = \frac{L}{h^2}\int_{x_{i-1}}^{x_{i+1}} K(S_h(x))\,dx + \int_{x_{i-1}}^{x_i} \phi(x)\frac{(x - x_{i-1})^2}{h^2}\,dx + \int_{x_i}^{x_{i+1}} \phi(x)\frac{(x - x_{i+1})^2}{h^2}\,dx \quad (38)$$

It is convenient to denote:

$$r_i^q = \frac{1}{h^{q+1}}\int_0^h \phi(x_i + y)y^q\,dy, \qquad i = 0 \dots n-1, \quad q = 1, 2 \quad (39)$$

$$l_i^q = \frac{1}{h^{q+1}}\int_{-h}^0 \phi(x_i + y)y^q\,dy, \qquad i = 1 \dots n, \quad q = 1, 2 \quad (40)$$

$$H(s) = \int_0^s K(x)\,dx \qquad \text{-monotone increasing function} \quad (41)$$

$\phi(x) \geq 0 \Rightarrow r_i^q \geq 0\ \forall q \in \mathbf{Z}_+$, $l_i^q \geq 0$ when q is even and $l_i^q \leq 0$ when q is odd.
Remark r_i^2 and l_i^2 give some kind of average of $\phi(x)/2$ to the right and left from x_i. The second element of (38) is equal to r_{i-1}^2,the third l_{i+1}^2.

We chose that S_h belongs to H_h^1. So S_h is linear on each segment $[x_i, x_{i+1}]$. For some constants a, b, c, d we can calculate integral

$$\int_a^b K(cx+d)\,dx = \frac{1}{c}\int_a^b K(cx+d)\,d(cx+d) = \frac{H(cb+d)-H(ca+d)}{c}$$

Let $c = (S_{i+1}-S_i)/h$, $d = S_i - x_i(S_{i+1}-S_i)/h$,

$$K_i = \frac{1}{h}\int_{x_i}^{x_{i+1}} K(S_h(x))\,dx$$

then

$$K_i = \begin{cases} K(S_i) & \text{if} \quad S_i = S_{i+1} \\ \frac{H(S_{i+1})-H(S_i)}{S_{i+1}-S_i} & \text{otherwise} \end{cases} \tag{42}$$

$$a_{ii} = \frac{L}{h}(K_{i-1}+K_i) + (r_{i-1}^2 + l_{i+1}^2)h, \qquad i = 1\ldots n-1 \tag{43}$$

$$a_{00} = \frac{L}{h}K_0 + l_1^2 h, \qquad a_{nn} = \frac{L}{h}K_{n-1} + r_{n-1}^2 h. \tag{44}$$

For other elements

$$a_{ii+1} = -\frac{L}{h^2}\int_{x_i}^{x_{i+1}} K(S_h(x))\,dx - \int_{x_i}^{x_{i+1}} \phi(x)\left[\frac{(x-x_{i+1})^2}{h^2} + \frac{x-x_{i+1}}{h}\right]dx$$

$$a_{ii+1} = -\frac{L}{h}K_i - (l_{i+1}^2 + l_{i+1}^1)h, \qquad i = 0\ldots n-1 \tag{45}$$

$$a_{ii-1} = -\frac{L}{h}K_{i-1} + \int_{x_{i-1}}^{x_i} \phi(x)\left[-\frac{(x-x_{i-1})^2}{h^2} + \frac{x-x_{i-1}}{h}\right]dx =$$

$$a_{ii-1} = -\frac{L}{h}K_{i-1} - (r_{i-1}^2 - r_{i-1}^1)h, \qquad i = 1\ldots n \tag{46}$$

$$b_i = \int_{x_{i-1}}^{x_{i+1}} K(S_h(x))P'(S_h(x))\frac{dS_h}{dx}\frac{d\psi_i}{dx}\,dx - \rho g\int_{x_{i-1}}^{x_{i+1}} K(S_h(x))\frac{d\psi_i}{dx}\,dx$$

Previous expression can be divided into four integrals: $I_1 + I_2 - \rho g(I_3 + I_4)$.

$$I_1 = \int_{x_{i-1}}^{x_i} K(S_h(x))P'(S_h(x))\frac{dS_h}{dx}\frac{1}{h}\,dx = \frac{1}{h}\int_{x_{i-1}}^{x_i} K(S_h)P'(S_h(x))dS_h(x) =$$

$$= \frac{G(S_h(x_i)) - G(S_h(x_{i-1}))}{h} = \frac{G(S_i) - G(S_{i-1})}{h}$$

where

$$G(s) = \int_0^s K(x)P'(x)dx \qquad \text{-monotone decreasing function} \tag{47}$$

Remark Linearity of $S_h(x)$ was not used.

$$I_2 = -\frac{G(S_{i+1}) - G(S_i)}{h}.$$

$$I_3 = \int\limits_{x_{i-1}}^{x_i} K(S_h(x))\frac{1}{h}\,dx = K_{i-1}, \qquad I_4 = -K_i$$

$$b_i = -\frac{G(S_{i+1}) - 2G(S_i) + G(S_{i-1})}{h} + \rho g(K_i - K_{i-1}) \tag{48}$$

$$b_0 = -\frac{G(S_1) - G(S_0)}{h} + \rho g K_0, \qquad b_n = \frac{G(S_n) - G(S_{n-1})}{h} - \rho g K_{n-1} \tag{49}$$

So we obtained expressions for a_{ij} and b_i: (43,44,45,46,48,49) with the help of (39,40,41,42,47). To show that it could be treated as a feasible approximation of (6) we divide a_{ij} and b_i by h and rewrite the liner equations in a form:

$$-\frac{1}{h}\left[K_i L\frac{u_{i+1} - u_i}{h} - K_{i-1}L\frac{u_i - u_{i-1}}{h}\right] + l^2_{i+1}(u_i - u_{i+1}) + r^2_{i-1}(u_i - u_{i-1}) +$$

$$+ r^1_{i-1}u_{i-1} - l^1_{i+1}u_{i+1} = -\left[\frac{G(S_{i+1}) - 2G(S_i) + G(S_{i-1})}{h^2} - \rho g\frac{K_i - K_{i-1}}{h}\right] \tag{50}$$

For $i = 1 \ldots n-1$. Additionally for $i = 0, n$:

$$-K_0 L\frac{u_1 - u_0}{h} + l^2_1(u_0 - u_1)h - l^1_1 u_1 h = -\frac{G(S_1) - G(S_0)}{h} + \rho g K_0 \tag{51}$$

$$K_{n-1}L\frac{u_n - u_{n-1}}{h} + r^2_{n-1}(u_n - u_{n-1})h + r^1_{n-1}u_{n-1}h = \frac{G(S_n) - G(S_{n-1})}{h} - \rho g K_{n-1} \tag{52}$$

Let us compare (50) with (6):

$$-\frac{1}{h}\left[K_i L\frac{u_{i+1} - u_i}{h} - K_{i-1}L\frac{u_i - u_{i-1}}{h}\right] \approx -\frac{\partial}{\partial x}\left[K(S)L\frac{\partial u}{\partial x}\right](x_i);$$

$$l^2_{i+1}(u_i - u_{i+1}) = l^2_{i+1}\frac{u_i - u_{i+1}}{h}h \approx l^2_{i+1}\frac{du}{dx}h \approx 0, \qquad r^2_{i-1}(u_i - u_{i-1}) \approx 0;$$

$$r^1_{i-1} = \frac{\phi_1}{h^2}\int_0^h y\,dy = \phi_1/2, \qquad l^1_{i+1} = -\phi_2/2$$

where $\phi_1 \in [\min\limits_{[x_{i-1},x_i]}\phi(x), \max\limits_{[x_{i-1},x_i]}\phi(x)]$, $\phi_2 \in [\min\limits_{[x_i,x_{i+1}]}\phi(x), \max\limits_{[x_i,x_{i+1}]}\phi(x)] \Rightarrow$

$$r^1_{i-1}u_{i-1} - l^1_{i+1}u_{i+1} = \frac{\phi_1 u_{i-1} + \phi_2 u_{i+1}}{2} \approx \phi(x_i)u(x_i);$$

$$-\left[\frac{G(S_{i+1}) - 2G(S_i) + G(S_{i-1})}{h^2}\right] \approx -\frac{d^2}{dx^2}G(S_h(x_i)) =$$

$$= -\frac{d}{dx}\left[G'(S_h(x_i))\frac{dS_h}{dx}(x_i)\right] = -\frac{d}{dx}\left[K(S_h(x_i))P'(S_h(x_i))\frac{dS_h}{dx}(x_i)\right];$$

$$\rho g\frac{K_i - K_{i-1}}{h} \approx \rho g\frac{d}{dx}K(S_h(x_i)).$$

Some remarks on implementation.
I1. Constants r_i^q, $i = 0 \dots n-1$, l_i^q, $i = 1 \dots n$, $q = 1, 2$ can be calculated once, a priory and then used many times.
I2 In general case, where $H(s)$, $G(s)$, $s \in [0,1]$ cannot be represented as a combination of elementary functions, we have a typical interpolation problem. A possible approach is to calculate $H(s_i)$, $G(s_i)$ in sufficiently many number of points with high accuracy by some quadrature formula for integrals; between them - interpolation by some simple function on each segment $[s_i, s_{i+1}]$. The simplest cases are linear interpolation:

$$H(s) \approx H(s_i) + \frac{H(s_{i+1}) - H(s_i)}{s_{i+1} - s_i}(s - s_i), \qquad s \in [s_i, s_{i+1}]$$

or the Taylor expansion: (s_i is the nearest point to s)

$$H(s) \approx H(s_i) + H'(s_i)(s - s_i) = H(s_i) + K(s_i)(s - s_i).$$

I3 The matrix $\{a_{ij}\}$ is tridiagonal, symmetric, positive definite, but may have no diagonal dominance. To solve it we can use, for instance, sweep method [7], p. 61 for diagonally dominant case or p. 86 for general case. Each calculation $A^h(S_h)$ needs $O(n)$ operations.

3.3 Another possible $A^h(S)$ by finite difference method.

In finite element method the basis equation was (8). Choosing finite difference method, we approximate the elliptic differential equation (6), where S is known and u is unknown.

Let grid be uniform, with step h.

$$\begin{array}{ll} B.C.1 & G_1 = \{x_i : \quad x_i = (i - 1/2)h, \quad i = 1 \dots n, \quad h = l/n\}, \\ B.C.2 & G_2 = \{x_i : \quad x_i = ih, \quad i = 0 \dots n, \quad h = l/n\}. \end{array} \tag{53}$$

$x_{i+1/2} = x_i + h/2$, $x_{i-1/2} = x_i - h/2$.

To obtain a difference scheme we use integro-interpolation method: we integrate the equation (6) on $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, for B.C.1 $i = 1 \dots n$ and $i = 1 \dots n-1$ for B.C.2:

$$\begin{aligned} &- \left[K(S(x))L\frac{du}{dx}(x)\right]_{x=x_{i-\frac{1}{2}}}^{x=x_{i+\frac{1}{2}}} + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \phi(x)u(x)\,dx = \\ &= - \left[K(S(x))\left(P'(S(x))\frac{dS}{dx}(x) - \rho g\right)\right]_{x=x_{i-\frac{1}{2}}}^{x=x_{i+\frac{1}{2}}} \end{aligned} \tag{54}$$

We have to approximate this equation using only values in grid points.

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \phi(x)u(x)\,dx \approx \phi_i u(x_i)h,$$

where $\phi_i = \phi(x_i)$ or

$$\phi_i = \frac{1}{h}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \phi(x)\,dx$$

It is also possible to use in the approximation neighbour values u_{i-1}, u_{i+1} (if they exists in Ω).

Other differential expressions are taken at the point $x_{i-\frac{1}{2}}$. To approximate them we will use values at x_{i-1} and x_i.
Case 1. $x_{i-1} = x_i - h \notin [0, l]$ is actual only for B.C.1, when $x_{i-\frac{1}{2}} = 0$. But in this case the flux is known at this point from the boundary condition (homogeneous case - flux f_l is zero):

$$\left[K(S)L\frac{du}{dx} - K(S)\left(P'(S)\frac{dS}{dx} - \rho g\right)\right]_{x=x_{i-\frac{1}{2}}=0} = f_l$$

Case 2. $x_{i-1} \in [0, l]$. For sufficiently smooth functions

$$\begin{array}{ll} K(S(x_{i-\frac{1}{2}})) = K(\frac{S_{i-1}+S_i}{2}) + O(h^2), & \frac{du}{dx}(x_{i-\frac{1}{2}}) = \frac{u_i - u_{i-1}}{h} + O(h^2) \\ P'(S(x_{i-\frac{1}{2}})) = P'(\frac{S_{i-1}+S_i}{2}) + O(h^2), & \frac{dS}{dx}(x_{i-\frac{1}{2}}) = \frac{S_i - S_{i-1}}{h} + O(h^2). \end{array} \tag{55}$$

For $x_{i+\frac{1}{2}}$ the situation is very similar.

Difference scheme: Let us substitute the continuous expressions in (54) by approximations from (55) and divide both sides by h. We will get a difference scheme in a following form:

$$-\frac{1}{h}\left[K_{i+\frac{1}{2}}L\frac{u_{i+1}-u_i}{h} - K_{i-\frac{1}{2}}L\frac{u_i - u_{i-1}}{h}\right] + \phi_i u_i =$$

$$= -\frac{1}{h}\left[K_{i+\frac{1}{2}}\left(P'_{i+\frac{1}{2}}\frac{S_{i+1}-S_i}{h} - \rho g\right) - K_{i-\frac{1}{2}}\left(P'_{i-\frac{1}{2}}\frac{S_i - S_{i-1}}{h} - \rho g\right)\right], \tag{56}$$

where

$$K_{i+\frac{1}{2}} = K\left(\frac{S_i + S_{i+1}}{2}\right), \qquad P'_{i+\frac{1}{2}} = P'\left(\frac{S_i + S_{i+1}}{2}\right)$$

$i = 2 \dots n-1$ for B.C.1 and $i = 1 \dots n-1$ for B.C.2.
Approximation of the boundary conditions:
B.C.1:

$$-K_{1+\frac{1}{2}}L\frac{u_2 - u_1}{h} + \phi_1 u_1 h = -f_l - K_{1+\frac{1}{2}}\left(P'_{1+\frac{1}{2}}\frac{S_2 - S_1}{h} - \rho g\right) \tag{57}$$

$$K_{n-\frac{1}{2}}L\frac{u_n - u_{n-1}}{h} + \phi_n u_n h = f_r - K_{n-\frac{1}{2}}\left(P'_{n-\frac{1}{2}}\frac{S_n - S_{n-1}}{h} - \rho g\right) \tag{58}$$

B.C.2:

$$u_0 = u_l, \qquad u_n = u_r; \qquad S_0 = S_l, \qquad S_n = S_r \tag{59}$$

(in homogeneous case $u_l = u_r = 0$). If $S(0,t) = S_l(t)$, $S(l,t) = S_r(t)$ are given then $u_l(t) = \frac{d}{dt}S_l(t)$, $u_r(t) = \frac{d}{dt}S_r(t)$.

Order of approximation
We are going to show that (56) approximates (6) and also (57),(58) or (59) approximate the corresponding boundary conditions with the second order.
$\circ$ Assume that functions u and S are sufficiently smooth and satisfy (6) and B.C.1 or B.C.2 (that can be non-homogeneous). Their Taylor expansions:

$$u_{i\pm1} = u_i \pm u'_i h + u''_i\frac{h^2}{2} \pm u'''_i\frac{h^3}{6} + O(h^4), \qquad S_{i\pm1} = S_i \pm S'_i h + S''_i\frac{h^2}{2} \pm S'''_i\frac{h^3}{6} + O(h^4).$$

Hence

$$\frac{S_{i\pm1}-S_i}{2} = \pm S_i'\frac{h}{2} + S_i''\frac{h^2}{4} + O(h^3),$$

$$\frac{u_{i\pm1}-u_i}{h} = \pm u_i' + u_i''\frac{h}{2} \pm u_i'''\frac{h^2}{6} + O(h^3) = u_+ \pm u_- + O(h^3),$$

$$K_{i\pm\frac{1}{2}} = K\left(\frac{S_{i\pm1}+S_i}{2}\right) = K\left(S_i + \frac{S_{i\pm1}-S_i}{2}\right) =$$

$$= K(S_i) + K'(S_i)\left(\pm S_i'\frac{h}{2} + S_i''\frac{h^2}{4}\right) + \frac{K''(S_i)(S_i'h)^2}{8} + O(h^3) = K_+ \pm K_- + O(h^3).$$

Using these formulas we can rewrite $[\dots]$ in the left side of (56):

$$\frac{1}{h}[\dots] = \frac{1}{h}\left[(K_+ + K_-)L(u_+ + u_-) - (K_+ - K_-)L(-u_+ + u_-) + O(h^3)\right]$$

$$= \frac{1}{h}[2K_+Lu_+ + 2K_-Lu_- + O(h^3)] =$$

where

$$K_+ = K(S_i) + K'(S_i)S_i''\tfrac{h^2}{4} + K''(S_i)(S_i')^2\tfrac{h^2}{8}, \quad u_+ = u_i''\tfrac{h}{2},$$
$$K_- = K'(S_i)S_i'\tfrac{h}{2}, \qquad\qquad\qquad\qquad u_- = u_i' + u_i'''\tfrac{h^2}{6}.$$

In $[\dots]$ we care only elements with order of h less than 3:

$$= \frac{1}{h}[K(S_i)Lu_i''h + K'(S_i)S_i'Lu_i'h + O(h^3)] = \frac{d}{dx}\left[K(S(x))L\frac{du}{dx}(x)\right]_{x=x_i} + O(h^2)$$

$\phi_i u_i = \phi(x)u(x)\mid_{x=x_i} + O(h^2)$ - depends on the choice of ϕ_i.

Now the right hand side. For the analogy with the left side $[\dots]$, we denote: $\tilde{K}(x) = K(x)P'(x)$. Then the Right Side of (56) is:

$$\text{R.S} = -\frac{1}{h}\left[\tilde{K}_{i+\frac{1}{2}}\frac{S_{i+1}-S_i}{h} - \tilde{K}_{i-\frac{1}{2}}\frac{S_i-S_{i-1}}{h}\right] + \rho g\frac{K_{i+\frac{1}{2}}-K_{i-\frac{1}{2}}}{h}$$

We already know the first element (we did very similar for u, K), for the second we use the Taylor expansion upstairs:

$$\frac{K_{i+\frac{1}{2}}-K_{i-\frac{1}{2}}}{h} = K'(S_i)S_i' + O(h^2) = \frac{d}{dx}K(S(x))\mid_{x=x_i} + O(h^2)$$

Returning from $\tilde{K}$ to K, P', we write the differential approximation

$$\text{R.S} = -\frac{d}{dx}\left[K(S(x_i))P'(S(x_i))\frac{dS}{dx}(x_i)\right] + \rho g\frac{d}{dx}K(S(x_i)) + O(h^2)$$

So the scheme (56) approximates (6) with second order. Now we investigate the boundary condition approximation.

B.C.2: (59) are exact, order: $O(h^k)$ for any k.

B.C.1: (58) is similar to (57), we will show the order of (57). u, S - known smooth functions, we can introduce a smooth flux function:

$$f(x) = K(S(x))L\frac{du}{dx}(x) - K(S(x))\left(P'(S(x))\frac{dS}{dx}(x) - \rho g\right)$$

Exact boundary condition – known flux: $f(0) = f_l$. We know that

$$K_{1+\frac{1}{2}} L \frac{u_2 - u_1}{h} - K_{1+\frac{1}{2}} \left(P'_{1+\frac{1}{2}} \frac{S_2 - S_1}{h} - \rho g \right) = f(x_{1+\frac{1}{2}}) + O(h^2)$$

We can rewrite (57) in a form:

$$f_l = -\phi_1 u_1 h + f(x_{1+\frac{1}{2}}) + O(h^2) \tag{57.1}$$

Using Taylor expansion: $f(x_{1+\frac{1}{2}}) = f(0) + f'(0)h + O(h^2)$

$$f_l = -\phi_1 u_1 h + f(0) + \frac{d}{dx} f(0) h + O(h^2) \tag{57.2}$$

In terms of fluxes equation (6) is: $\phi(x)u(x) = \frac{d}{dx} f(x)$. We cannot use it at the point $x = 0$ but it is true for $x = 0+$, all positive points from some small neighbourhood of 0.

$$f_l = -\phi_1 u_1 h + f(0) + \phi(0) u(0) h + O(h^2) \tag{57.3}$$

$\phi_1 u_1 - \phi(0)u(0) = O(h)$, finally we have the second order of approximation:

$$f_l = f(0) + O(h^2) \tag{57.4}$$

• One can expect a second order of convergence.

On solvability and implementation

Obtained SLAE has tridiagonal matrix $\{a_{ij}\}$ with diagonal domination:

$$a_{ii} = \frac{1}{h^2}(K_{1+\frac{1}{2}} + K_{1-\frac{1}{2}}) + \phi_i = |a_{ii-1}| + |a_{ii+1}| + \phi_i > |a_{ii-1}| + |a_{ii+1}|$$

also for boundary conditions: B.C.1 $a_{11} > |a_{12}|$, $a_{nn} > |a_{nn-1}|$;
B.C.2 $a_{00} = 1 > |a_{01}| = 0$, $a_{nn} = 1 > |a_{nn-1}| = 0$. The sweep method ([7], p. 61) can be used to solve the system. It needs $O(n)$ operations.

We can combine this finite difference scheme or finite element method from previous section or some other method for elliptic problem (6) with some numerical method for Ordinary Differential Equation. The interesting question is what resulting order will have this combination. Suppose the method in "x direction" gives $O(h^q)$ error and the method in "t direction" is of p-th order. Our hypothesis is that the combination may have $O(h^q + \Delta t^p)$ error.

We combined the finite difference method discussed above with Euler, Improved Euler, Cauchy-Euler and Fourth Order Runge-Kutta methods for ODE. Next section we report about results obtained in computational experiments.

4 Computational Experiment

To solve Initial Value Problem (4), we implemented:
in "x direction" the finite difference method (56) with (57),(58) and (59);
in "t direction" Runge-Kutta type methods (29),(30), (31),(32). We used these numerical methods for solving several IVP with homogeneous boundary conditions B.C.1 or B.C.2 (constant values S_l, S_r):
IVP1. $l = 0.1$; B.C.1; Initial value:

$$S(0,x) = 2\left(\frac{x - l/2}{l}\right) + 0.1$$

IVP2. $l = 0.1$; B.C.2; Initial value – the same like in **IVP1**
IVP3. $l = 0.1$; B.C.1;

$$S(0,x) = 0.5 \exp\left\{-100\left(\frac{x - \frac{2}{3}l}{l}\right)^2\right\}$$

IVP4. $l = 0.1$; B.C.2; Initial value – the same like in **IVP3**
We remark that initial values for **IVP1** – **IVP4** are smooth and bounded away from 0 and 1.
IVP5. $l = 0.1$; B.C.2: $S(t,0) = 1$, $S(t,l) = 0.5$; Initial value:

$$S(0,x) = \begin{cases} 0.2 & x \in [0, l/4] \\ 0.5 & x \in [l/4, l/2] \\ 2|x - 3l/4| & x \in [l/2, l] \end{cases}$$

This problem has difficulties: discontinuity of initial value, the function reaches zero at the point $x = 3l/4$.

Other functions and parameters that were used:

$$K(S) = K_0 S^3, \quad \text{with} \quad K_0 = 0.015,$$

$$P(S) = \frac{p_1}{2^3} - p_1(S - 1/2)^3 - p_2(S - 1), \quad \text{with} \quad p_1 = 5.0, \quad p_2 = 0.1$$

$\phi \in (0,1)$ – some constant, doesn't depend on x. (In spite of non-physical meaning we set $\phi = 1$. For other ϕ we can divide the function K on it, $K := K/\phi$).

Each problem **IVP1** – **IVP5** we calculated using the discussed methods for several positive constants $L \in L^* = \{0.0001, 0.001, 0.01, 0.1, 1.0\}$. The reason for such choice was not only the comparison of numerical solutions for different L, but also we wanted to compare them with the "limit case" – Richard's equation. The well-known Richard's equation

$$\phi \frac{\partial S}{\partial t} = -\frac{\partial}{\partial x} F(S) = -\frac{\partial}{\partial x}\left[K(S)\left(\frac{\partial P(S)}{\partial x} - \rho g\right)\right] \tag{60}$$

can be obtained from (4) by setting formally $L = 0$. A numerical algorithm that was used for Richard's equation we describe in the next section. We remark that the condition $L > 0$ is important for all methods that we implemented for (4), and for Richard's equation we have to use another method.

Figures 1-5 present saturation profiles $S(t_k, x)$ at selected time moments t_k for **IVP1** – **IVP5**. Each color corresponds to some value of parameter L:

red – $L = 0$, blue – $L = 0.0001$, green – $L = 0.001$,
magenta – $L = 0.01$, cyan – $L = 0.1$, black – $L = 1.0$.

Each line has a corresponding number (it is near the line and has the same color) which means t_k. It was convenient to divide results for some IVP into two pictures. The upper contains results for $L = 0$ and the smallest L: $L = 0.0001$, $L = 0.001$. Red lines were plotted first, then blue and green lines. For **IVP1** – **IVP4** they are rather closed to each other so the red was covered by blue first and then the green covers red and blue. In this case time t_k is printed once in red for red, blue and green lines. The initial value $S(0, x)$ is the same for all L. It was plotted in the upper picture with red color. In the lower picture it is possible to see the difference between different $L > 0$. For each L it has 3-4 profiles. Next we give the content of figures more precise with order in which lines were plotted.

IVP1 Figure 1
Upper picture:

red(0.0) $t \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 2.0, 3.0, 4.0, 5.0, 10.0\}$
blue(0.0001) $t \in \{0.2, 0.4, 0.6, 0.8, 1.0, 2.0, 3.0, 4.0, 5.0, 10.0\}$
green(0.001) $t \in \{0.2, 0.4, 0.6, 0.8, 1.0, 2.0, 3.0, 4.0, 5.0, 10.0\}$

Lower picture:

1)black(1.0) $t \in \{0.2, 1.0, 5.0\}$, 2)magenta(0.01) $t \in \{0.2, 1.0, 5.0\}$,
3)cyan(0.1) $t \in \{0.2, 1.0, 5.0\}$, 4)green(0.001) $t \in \{0.2, 1.0, 5.0\}$,
5)blue(0.0001) $t \in \{0.2, 1.0, 5.0\}$.

IVP2 Figure 2
Upper picture:

red(0.0) $t \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.5\}$
blue(0.0001) $t \in \{0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.5\}$
green(0.001) $t \in \{0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.5\}$
magenta(0.01) $t \in \{0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.5\}$

Lower picture:

1)black(1.0) $t \in \{1.0, 2.0, 3.0, 5.0, 10.0\}$,
2)cyan(0.1) $t \in \{0.6, 1.0, 1.4, 2.0\}$, 3)magenta(0.01) $t \in \{0.2, 0.6, 1.0\}$.

IVP3 Figure 3
Upper picture:

red(0.0) $t \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 2.0, 3.0, 4.0, 5.0, 10.0\}$
blue(0.0001) $t \in \{0.2, 0.4, 0.6, 0.8, 1.0, 2.0, 3.0, 4.0, 5.0, 10.0\}$
green(0.001) $t \in \{0.2, 0.4, 0.6, 0.8, 1.0, 2.0, 3.0, 4.0, 5.0, 10.0\}$

Lower picture:

1)black(1.0) $t \in \{0.2, 1.0, 5.0\}$, 2)cyan(0.1) $t \in \{0.2, 1.0, 5.0\}$,
3)magenta(0.01) $t \in \{0.2, 1.0, 5.0\}$, 4)green(0.001) $t \in \{0.2, 1.0, 5.0\}$,
5)blue(0.0001) $t \in \{0.2, 1.0, 5.0\}$.

IVP4 Figure 4
Upper picture:

red(0.0)	$t \in \{0.0, 0.2, 0.4, 0.6, 1.2, 2.0, 5.0, 10.0\}$
blue(0.0001)	$t \in \{0.2, 0.4, 0.6, 1.2, 2.0, 5.0, 10.0\}$
green(0.001)	$t \in \{0.2, 0.4, 0.6, 1.2, 2.0, 5.0, 10.0\}$

Lower picture:

1)black(1.0) $t \in \{0.2, 1.2, 5.0\}$, 2)cyan(0.1) $t \in \{0.2, 1.2, 5.0\}$,
3)magenta(0.01) $t \in \{0.2, 1.2, 5.0\}$, 4)green(0.001) $t \in \{0.2, 1.2, 5.0\}$,
5)blue(0.0001) $t \in \{0.2, 1.2, 5.0\}$.

IVP5 Figure 5
Upper picture:

red(0.0)	$t \in \{0.01, 0.03, 0.05, 0.1, 0.2, 0.4, 0.6, 1.0, 2.0\}$
blue(0.0001)	$t \in \{0.01, 0.03, 0.05, 0.1, 0.2, 0.4, 0.6, 1.0, 2.0\}$
green(0.001)	$t \in \{0.01, 0.03, 0.05, 0.1, 0.2, 0.4, 0.6, 1.0, 2.0\}$
red(0.0)	$t \in \{0.0\}$

Lower picture:

1)green(0.001) $t \in \{0.01, 0.5, 2.0\}$, 2)black(1.0) $t \in \{1.0, 2.0, 10.0\}$,
3)cyan(0.1) $t \in \{0.2, 1.0, 3.0\}$, 4)magenta(0.01) $t \in \{0.04, 0.08, 0.5, 2.0\}$.

Figures 1,2 were obtained by Cauchy-Euler, figures 3,4 by Improved Euler methods with 270 grid points in $[0, l]$ for B.C.1 and 240 grid points for B.C.2. To obtain Figure 5 we used 480 points in $[0, l]$ and Cauchy-Euler method. Red lines (numerical solution for the Richard's equation) were obtained by following method with 200 points in x. Time steps Δt were different for different L. We remark only that for smaller $L > 0$ we need smaller time step (for example from green to blue – approximately 10 times smaller), when the calculated profile is non-smooth with high frequency oscillations then decreasing Δt may change the situation to more regular.

4.1 Algorithm for Richard's equation

For the Richard's equation we use the same grid as for finite difference method from previous section (53). Boundary conditions:
B.C.1: $F(S(0)) = 0$, $F(S(l)) = 0$; ($L = 0$ also in boundary conditions)
B.C.2: $S(t, 0) = S_l$, $S(t, l) = S_r$.
Integrating over $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [t^j, t^{j+1}]$ (60) and dividing by $h\Delta t$ we can get a balance equation:

$$\frac{1}{h}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \phi \frac{S(t^{j+1}) - S(t^j)}{\Delta t}\, dx = -\frac{1}{\Delta t}\int_{t^j}^{t^{j+1}} \frac{1}{h} K(S)\left[\frac{\partial P(S)}{\partial x} - \rho g\right]\Bigg|_{x_{i-1/2}}^{x_{i+1/2}} dt$$

We approximate integrals using quadrature formulas: left integral - a central point formula, right integral - two points formula at t_j, t_{j+1} with weight $\alpha = 0.5$:

$$\phi_i \frac{S_i^{j+1} - S_i^j}{\Delta t} = -\left[\alpha \frac{F_{i+1/2}^{j+1} - F_{i-1/2}^{j+1}}{h} + (1-\alpha)\frac{F_{i+1/2}^{j} - F_{i-1/2}^{j}}{h}\right], \quad x_i \in (0, l), \tag{61}$$

only for B.C.2 we additionally have $S_0 = S_l$, $S_n = S_r$,

$$F^k_{i+1/2} = \begin{cases} K\left(\frac{S^k_i + S^k_{i+1}}{2}\right)\left[\frac{P(S^k_{i+1}) - P(S^k_i)}{h} - \rho g\right], & \text{if} \quad x_{i+1/2} \in (0,l) \\ 0 & \text{if} \quad x_{i+1/2} \in \partial(0,l) \end{cases}$$

The second condition is never satisfied with B.C.2. In B.C.1 case we have zero flux boundary condition at this point, that's why it is zero. This scheme is implicit. To obtain an approximate solution at a new time step S^{j+1} we need to solve a nonlinear algebraic system of equations with number of unknowns equal to number of equations. We can rewrite (61) in more convenient for iterations form:

$$\begin{array}{l} S^{j+1}_i = S^j_i + \frac{\Delta t}{h}\left[\alpha \Delta F_i(S^{j+1}) + (1-\alpha)\Delta F_i(S^j)\right], \quad x_i \in (0,l), \\ S^{j+1}_0 = S_l, \quad S^{j+1}_n = S_r \quad \text{additionally for B.C.2} \end{array} \tag{62}$$

where

$$\Delta F_i(S^k) = (F^k_{i+1/2} - F^k_{i-1/2})$$

In Vector form (62) looks:

$$S^{j+1} = R(S^{j+1}, S^j)$$

We used the following iteration process to find S^{j+1}:

$$S^{(0)} = S^j, \quad \ldots \quad , S^{(p+1)} = R(S^{(p)}, S^j),$$

until $\|S^{(p+1)} - S^{(p)}\|_\infty < \epsilon$ at some $p \le P \Rightarrow S^{j+1} := S^{(p+1)}$
otherwise ($\|S^{(P)} - S^{(P-1)}\|_\infty \ge \epsilon$) reduce Δt and begin the iteration process again using the same S^j and new Δt.

4.2 Comparison of results obtained on different grids

The second test that we can apply to the numerical methods is a numerical estimation of convergence order. We use here some simple variant of Richardson's extrapolation method (see for example [8], Ch 6, p. 267). Let us consider a sequence of uniform grids (53) embedded one to another

$$G^1 \subset G^2 \subset \ldots \subset G^I$$

such that corresponding parameter h_i (distance between nodes) decreases in k times from grid G^i to G^{i+1}: $h_i = kh_{i+1}$, $k \in \mathbf{Z}$. Assume that we have some numerical method with supposed order of convergence $O(h^q + \Delta t^p)$. Choosing some time step Δt_i we can calculate an approximate solution S_i at the points of time grid $\mathcal{T}_i \subset [0,T]$ for each grid G^i using this method. It is convenient to choose Δt_i, $\mathcal{T}_i$ in such way that exists not empty intersection $\mathcal{T} = \bigcap_i \mathcal{T}_i$. Then we can compare solutions S_i at points $(x,t) \in G^1 \times \mathcal{T}$ (since $G^1 \times \mathcal{T} \subset G^i \times \mathcal{T}_i$). Convergence of the method means that $|S_i(x,t) - S(x,t)|$ for $(x,t) \in G^1 \times \mathcal{T}$ becomes smaller when i increases. When we have order (q,p) then

$$E_i(t) = \|S_i(x,t) - S(x,t)\|_{G^1} \sim C_1 h_i^q + C_2 \Delta t_i^p,$$

where S is exact solution. Assuming that $k^q(\Delta t_{i+1})^p \approx (\Delta t_i)^p$ and doing formally the same procedure for $i+1$ we will get:

$$E_{i+1}(t) = \|S_{i+1}(x,t) - S(x,t)\|_{G^1} \sim \frac{1}{k^q}(C_1 h_i^q + C_2 \Delta t_i^p) \sim \frac{1}{k^q} E_i(t). \tag{63}$$

We don't know $S(x,t)$, on $G^1 \times \mathcal{T}$ but we can substitute S by S^I – a numerical solution obtained on the finest grid G^I. So let $S := S^I$ and we can calculate $E_i(t)$ at all points $t \in \mathcal{T}$ for $i = 1 \ldots I-1$ From (63) we have approximate equalities:

$$E_i(t)/E_{i+1}(t) \approx k^q \tag{64}$$

We use two $E_i(t)$ due to different norms:

$$E_i^\infty(t) = \max_{x \in G^1} |S_i(x,t) - S_I(x,t)| \quad \text{or} \quad E_i^1(t) = \frac{1}{|G^1|} \sum_{x \in G^1} |S_i(x,t) - S_I(x,t)|.$$

We expect that having a correct supposition about order (q,p) and using the method for solving different "smooth" problems we will have E_i/E_{i+1} around k^q. This procedure proves nothing it can only give an assurance that the predicted order is right (if (64) is rather accurate) or wrong (if (64) is "too bad").

In our case we used
G^1–G^5 ($I = 5$, $k = 3$) for B.C.1 with $\{30, 90, 270, 810, 2430\}$ points in $[0, l]$,
G^1–G^7 ($I = 7$, $k = 2$) for B.C.2 with $\{30, 60, 120, 240, 480, 960, 1920\}$ points in $[0, l]$.
The grid $\mathcal{T}$ in "t direction" consists of several points (usually 9) from the time interval where the solution essentially changes. Mostly 5 first points were with small intervals between them and then, additionally, 4 points with intervals 5 times larger. We have four methods in "t" direction: Euler [E] (29), Improved Euler [IE](30), Cauchy-Euler [CE](31) and Runge-Kutta 4-th order [RK4](32). Our hypotheses are that these methods have order (q,p): $(2,1)$, $(2,2)$, $(2,2)$, $(2,4)$.
$(q,p) = (2,2)$: $h_{i+1} = h_i/k$, $\Delta t_{i+1} = \Delta t_i/k$;
$(q,p) = (2,4)$: $h_{i+1} = h_i/k$, $\Delta t_{i+1} \approx \Delta t_i/\sqrt{k}$ (to reach exactly $t^{j+1} \in \mathcal{T}$ from $t^j \in \mathcal{T}$ we can determine Δt_{i+1} from

$$\Delta t_{i+1} = (t^{j+1} - t^j) \Big/ \left(\left\downarrow \sqrt{k} \frac{t^{j+1} - t^j}{\Delta t_i} \right\downarrow + 1 \right) \approx \Delta t_i/\sqrt{k}$$

$\downarrow \cdot \downarrow$ - the nearest smallest integer.

We did the test described before for the following cases
IVP1, **IVP2** were calculated by CE method for L from L^*;
IVP3, **IVP4** were calculated by IE method for L from L^*;
IVP3, **IVP4** were calculated by RK4 method for $L \in \{0.0001, 0.001, 0.01\}$;
Next there are typical results that we obtained. G^1 contains 30 points, $|G^1| = 30$. Each row corresponds to some time moment $t \in \mathcal{T}$. Each column has some number i and contains comparison S_i with S_{i+1}

$$\begin{array}{llll} (30/90 \sim i = 1, & 90/270 \sim i = 2, & 270/810 \sim i = 3; & 30/60 \sim i = 1, \\ 60/120 \sim i = 2, & 120/240 \sim i = 3, & 240/480 \sim i = 4, & 480/960 \sim i = 5). \end{array}$$

An element of the tables for some $t \in \mathcal{T}$, $i \in \{1, \ldots I-2\}$ is a pair

$$\left[E_i^\infty(t)/E_{i+1}^\infty(t)\,,\; E_i^1(t)/E_{i+1}^1(t) \right].$$

The information attached to each table contains the problem's label, the method, L, boundary condition type, total number of grids I, expected value k^2 and two numbers

$$\max_{t \in \mathcal{T}} E_{I-1}^\infty(t), \qquad \max_{t \in \mathcal{T}} E_{I-1}^1(t).$$

time	30/90		90/270		270/810	
0.2	14.4	13.1	10.7	9.9	10.1	10.0
0.4	16.4	13.3	8.8	8.9	10.1	10.0
0.6	12.6	11.1	10.0	9.2	10.1	10.1
0.8	10.8	9.6	10.9	9.6	10.1	10.1
1.0	9.8	8.6	11.5	10.0	10.1	10.1
2.0	6.8	9.2	10.7	10.0	10.1	10.1
3.0	7.1	8.8	11.0	10.2	10.1	10.1
4.0	6.4	6.7	10.2	9.8	10.4	10.1
5.0	15.2	16.4	8.8	9.2	10.1	10.1

IVP1
Cauchy-Euler
L=0.0001
B.C.1
$I = 5$, $k^2 = 9$
4.857916e-05
5.560761e-06

time	30/60		60/120		120/240		240/480		480/960	
0.1	3.3	3.2	4.1	4.2	4.2	4.1	4.2	4.2	5.0	5.0
0.2	5.3	5.4	3.7	3.8	4.1	4.1	4.2	4.2	5.0	5.0
0.3	5.6	4.6	3.6	3.9	4.0	4.1	4.2	4.2	5.0	5.0
0.4	5.4	4.5	3.7	3.9	4.0	4.1	4.2	4.2	5.0	5.0
0.5	5.2	4.6	3.8	4.0	4.0	4.1	4.2	4.2	5.0	5.0
0.6	5.0	4.5	3.9	4.0	4.0	4.1	4.2	4.2	5.0	5.0
0.7	4.8	4.6	4.0	4.1	4.1	4.1	4.2	4.2	5.0	5.0
0.8	4.7	4.7	4.0	4.1	4.1	4.1	4.2	4.2	5.0	5.0
0.9	4.7	4.9	4.1	4.2	4.1	4.1	4.2	4.2	5.0	5.0
1.0	5.0	5.4	4.0	4.1	4.1	4.1	4.2	4.2	5.0	5.0
1.1	5.4	5.4	4.2	4.2	4.1	4.1	4.2	4.2	5.0	5.0
1.2	9.5	6.9	4.3	4.3	4.1	4.1	4.2	4.2	5.0	5.0

IVP2
Cauchy-Euler
L=0.0001
B.C.2
$I = 7$, $k^2 = 4$
3.685478e-05
4.675827e-06

time	30/90		90/270		270/810	
0.2	8.4	8.9	9.1	9.1	10.0	10.0
0.4	9.4	8.7	9.1	9.1	10.0	10.0
0.6	9.4	8.9	9.1	9.1	10.0	10.0
0.8	9.3	9.0	9.1	9.1	10.0	10.0
1.0	8.8	9.0	9.1	9.1	10.0	10.0
2.0	9.3	9.2	9.1	9.1	10.0	10.0
3.0	30.0	19.9	14.3	9.7	10.2	10.0
4.0	15.4	17.1	8.8	9.1	10.0	10.0
5.0	5.9	5.6	10.4	10.5	10.0	10.1

IVP3
Runge-Kutta 4
L=0.01
B.C.1
$I = 5$, $k^2 = 9$
5.262730e-05
3.144243e-06

time	30/60		60/120		120/240		240/480		480/960	
0.2	4.3	3.9	4.1	4.0	4.1	4.0	4.2	4.2	5.0	5.0
0.4	3.9	4.1	4.0	4.0	4.0	4.1	4.2	4.2	5.0	5.0
0.6	4.0	4.0	4.0	4.0	4.1	4.0	4.2	4.2	5.0	5.0
0.8	3.9	4.1	4.0	4.0	4.0	4.0	4.2	4.2	5.0	5.0
1.0	4.1	4.0	4.0	4.0	4.1	4.1	4.2	4.2	5.0	5.0
2.0	3.6	3.9	3.8	4.0	4.0	4.0	4.2	4.2	5.0	5.0
3.0	3.8	3.9	3.9	3.9	4.0	4.0	4.2	4.2	5.0	5.0
4.0	3.9	3.9	3.9	4.0	4.0	4.0	4.2	4.2	5.0	5.0
5.0	4.0	4.0	4.0	4.0	4.0	4.0	4.2	4.2	5.0	5.0

IVP4
Runge-Kutta 4
L=0.01
B.C.2
$I = 7$, $k^2 = 4$
5.275495e-06
9.851936e-07

time	30/90		90/270		270/810	
0.2	8.9	8.9	9.1	9.1	10.0	10.0
0.4	8.8	8.9	9.1	9.1	10.0	10.0
0.6	8.8	8.9	9.1	9.1	10.0	10.0
0.8	8.8	8.9	9.1	9.1	10.0	10.0
1.0	8.8	8.8	9.1	9.1	10.0	10.0
2.0	8.7	8.8	9.1	9.1	10.0	10.0
3.0	8.9	8.9	9.1	9.1	10.0	10.0
4.0	8.9	8.8	9.1	9.1	10.0	10.0
5.0	8.9	8.8	9.1	9.1	10.0	10.0

IVP3
Improved Euler
L=1.0
B.C.1
$I = 5$, $k^2 = 9$
2.234108e-06
8.420502e-07

time	30/60		60/120		120/240		240/480		480/960	
0.2	4.0	4.0	4.0	4.0	4.0	4.0	4.2	4.2	5.0	5.0
0.4	3.9	4.0	4.0	4.0	4.0	4.0	4.2	4.2	5.0	5.0
0.6	3.9	4.0	4.0	4.0	4.0	4.0	4.2	4.2	5.0	5.0
0.8	3.9	4.0	4.0	4.0	4.0	4.0	4.2	4.2	5.0	5.0
1.0	3.9	4.0	4.0	4.0	4.0	4.0	4.2	4.2	5.0	5.0
2.0	3.9	3.9	4.0	4.0	4.0	4.0	4.2	4.2	5.0	5.0
3.0	3.9	4.0	4.0	4.0	4.0	4.0	4.2	4.2	5.0	5.0
4.0	3.8	3.9	3.9	4.0	4.0	4.0	4.2	4.2	5.0	5.0
5.0	3.7	3.8	3.8	3.9	4.0	4.0	4.2	4.2	5.0	5.0

IVP4
Improved Euler
L=1.0
B.C.2
$I = 7$, $k^2 = 4$
1.359211e-06
5.124769e-07

The numerical data seem to be consistent with expected value $k^2 \in \{9, 4\}$ And the finite difference method (56),(57), (58),(59) combined with IE, CE, RK4 has a good chance to be of order $O(h^2 + \Delta t^2)$, $O(h^2 + \Delta t^2)$, $O(h^2 + \Delta t^4)$.

Remarks The initial Δt_1 was chosen sufficiently small to have a "smooth" numerical solutions for all $i = 1 \dots I$.

In **IVP5** we have an opposite situation. It seems that the methods don't have high order for this problem, which has discontinuous initial value.

All algorithms were implemented in ANSI C code. Pictures were printed using MATLAB.

5 Conclusion

We consider initial-boundary value problem with modified Richard"s equation suggested in [1], [9] in one dimensional spatial case and two types of boundary conditions: zero flux or given, constant in time values on the boundary. This problem was transformed into another form, namely initial value problem with Abstract Ordinary Differential Equation, using a solvability of elliptic boundary value problem (weak solvability by variational approach). We used deep similarity of this form with ODE in $\mathbf{R}$ to show local existence, uniqueness and continuous dependence on initial data provided the initial data is bounded away from zero. Like in ODE case a local solution can be extended at least while it is bounded away from zero and one. This solution being an abstract function on x at each time moment t is also a continuous function on variables x and t and has a continuous partial derivative in t.

We have also used the AODE form in order to obtain numerical methods for modified Richard's equation: we combined some numerical method for elliptic boundary-value problem ("x direction") with some numerical method for ODE ("t direction") and the choice of the methods can be rather independent from each other. For elliptic problem we described in details two possibilities by finite element approach and finite difference approach. In "t direction" we can choose for example explicit Runge-Kutta type methods like Euler, Improved Euler, Cauchy-Euler and Runge-Kutta 4th order methods. Finite element method is a natural numerical method for variational approach that was used in AODE form. We used this connection to show the convergence of Finite element $-$ Euler method.

Four combinations, namely the finite difference method (that has a second order of approximation) in "x direction" with four already mentioned methods in "t direction", were implemented in ANSI C program. Numerical results, obtained on the nested sequence of grids for smooth initial functions bounded away from zero, have showed that the last three methods have second order in "x" and 2nd, 2nd, 4th orders in "t" respectively.

The modified Richard's equation with strictly positive constant parameter L has one additional term, the highest derivative has one order more comparing with standard Richard's equation. If $L = 0$, formally, then the first equation becomes equal to the second. The mentioned numerical methods are undefined for this case, but for small positive L (on the same "good" initial functions used in the previous test) the obtained results were rather closed to the numerical results obtained for Richard's equation. In some sense the modified Richard's equation includes the standard Richard's equation. Another aspect for comparison of these two equations can be the similarity of the form in which the Richard's equation is written with AODE form: time derivative in the left hand side and an operator that acts on functions as they would have had only dependence on x in the right hand side. The right hand side operator is more regular for the AODE form.

We used the restricted case in order to have more simple situation. It may be possible to generalize the problem to a more complicated one with non-homogeneous boundary conditions, coefficients depending on more variables or even introduce additional terms and consider multidimensional case.

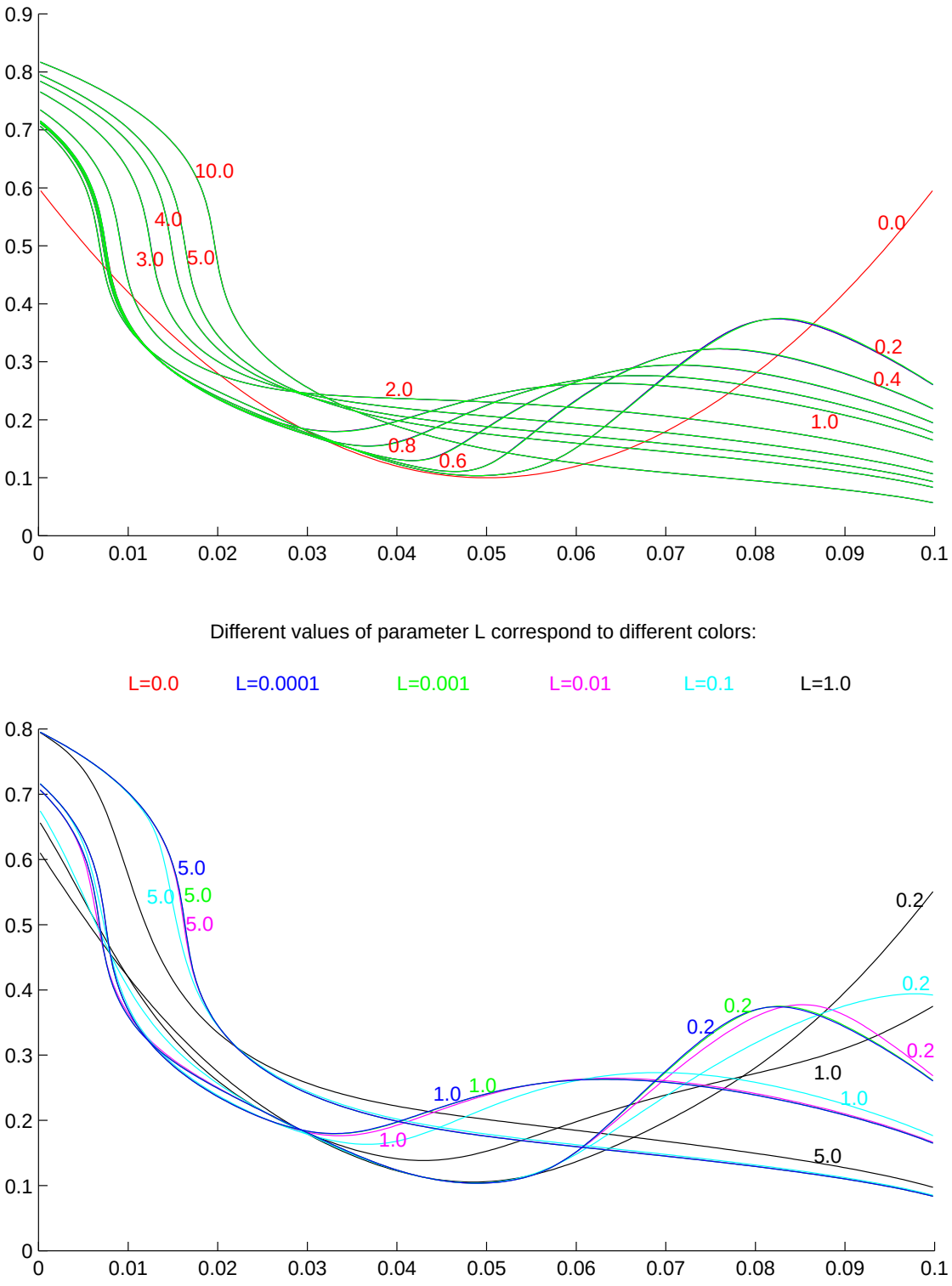


Figure 1: IVP1

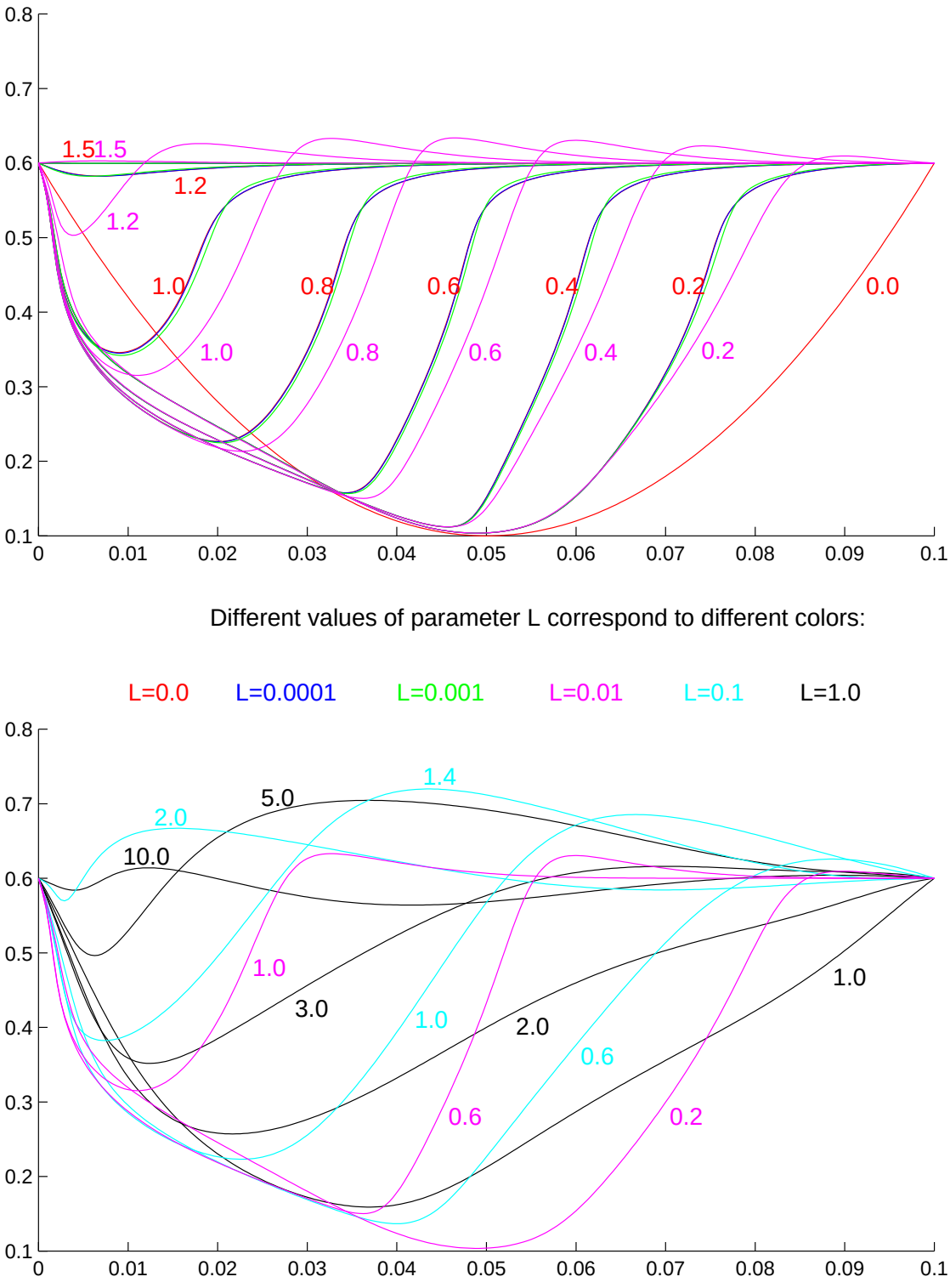


Figure 2: IVP2

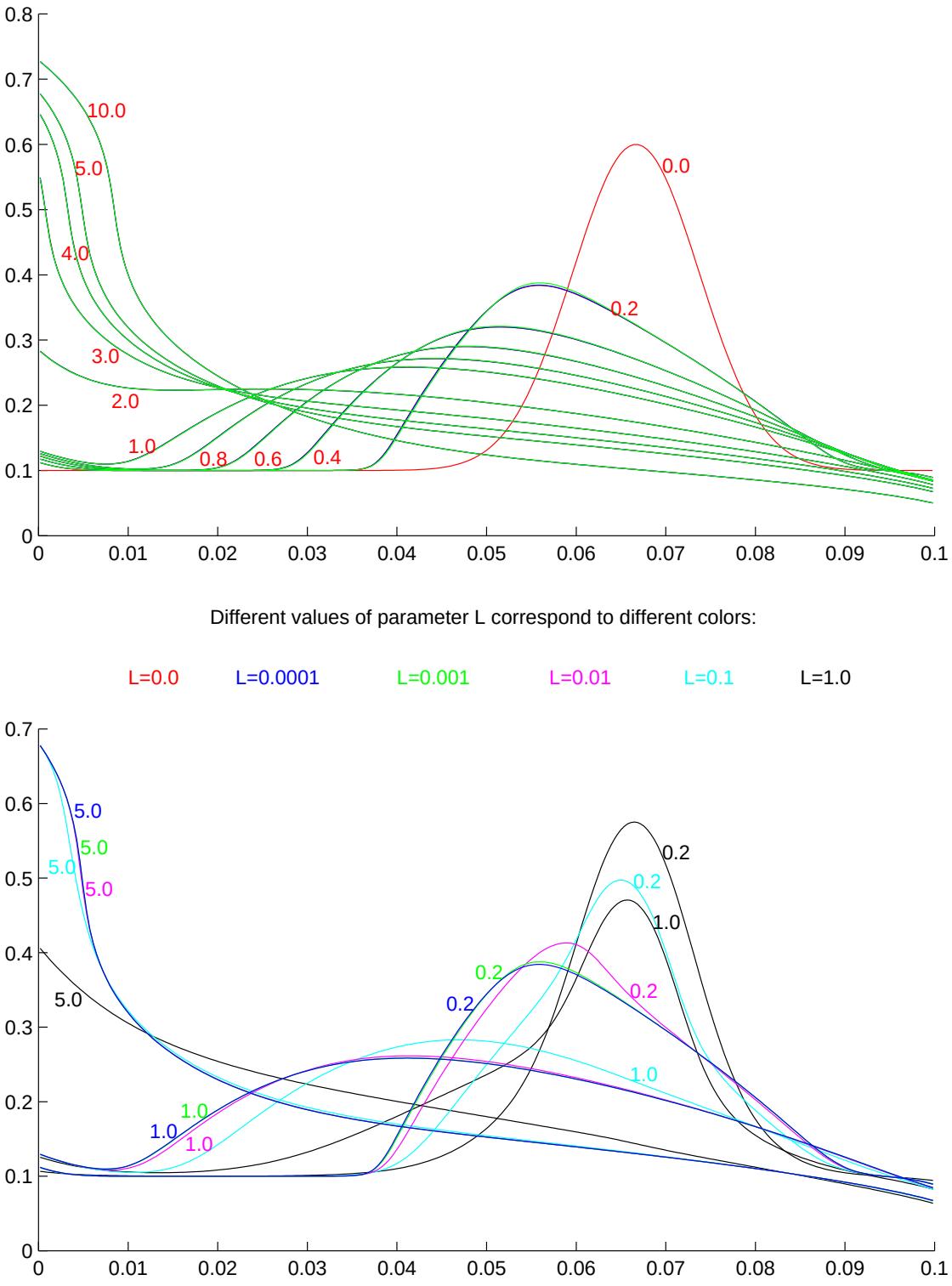
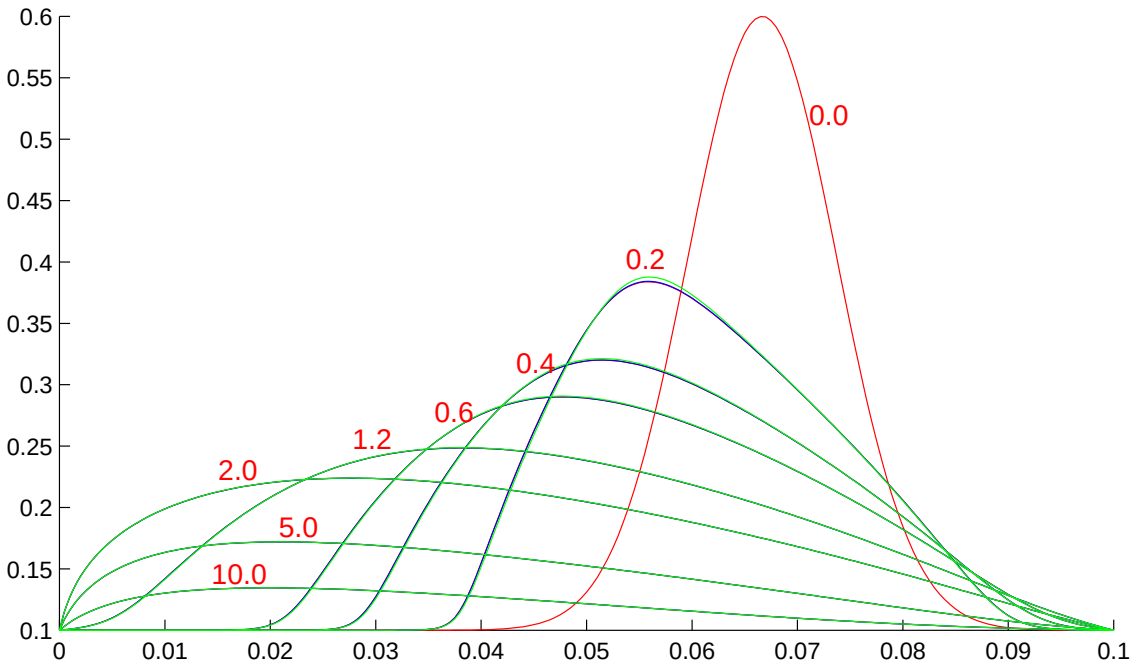


Figure 3: IVP3



Different values of parameter L correspond to different colors:

L=0.0 L=0.0001 L=0.001 L=0.01 L=0.1 L=1.0

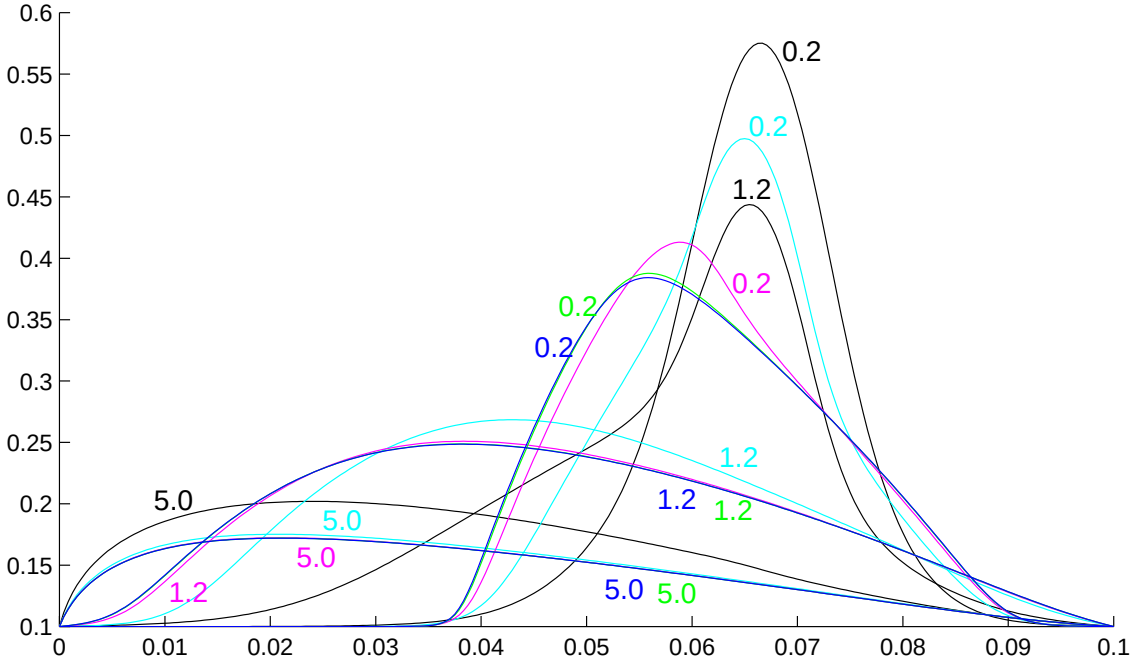


Figure 4: IVP4

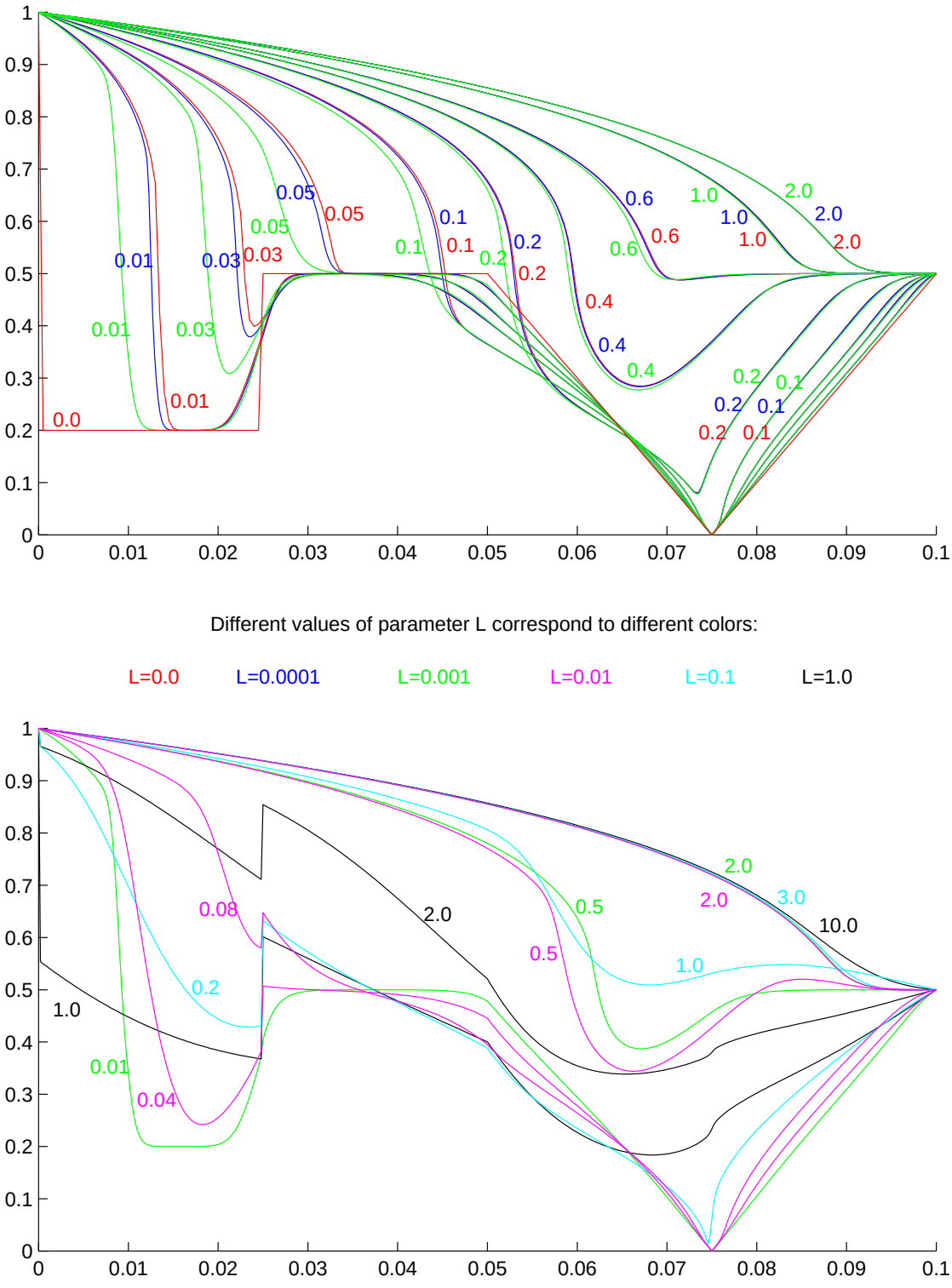


Figure 5: IVP5

References

[1] J. Hulshof, J. R. King, "Analysis of a Darcy Flow Model with a Dynamic Pressure Saturation Relation", SIAM J. APPL. MATH. Vol. 59, No 1. pp 318-346.

[2] R. A. Adams, "Sobolev Spaces", Academic Press, New York, 1975.

[3] M. Junk, "Analytical and Numerical Methods for Elliptic Partial Differential Equations", http://www.mathematik.uni-kl.de/∼junk/
a) part "General Dirichlet Problem"
b) part "Finite Element Method"

[4] L. A. Lusternik, V. J. Sobolev, "Elements of Functional Analysis", Chapter 6, Gordon and Breach, science publishers, 1968.

[5] D. Braess, "Finite Elemente", Springer 1992.

[6] F. Verhulst, "Nonlinear Differential Equations and Dynamical Systems", Springer 1996.

[7] A.A. Samarskii, E.S. Nikolaev "Numerical Methods for Grid Equations", Vol 1, Birkhäuser Verlag, 1989.

[8] G.I. Marchuk "Methods of Numerical mathematics", Springer Verlag, 1982.

[9] S. M. Hassanizadeh, W.G. Gray "Thermodynamic basis of capillary pressure in porouse media", Water Resources Research, 29 (1993) pp3389-3405.

[10] F. Stauffer "Time dependence of the relations between capillary pressure, water content and conductivity during drainage of porous media" in International Association of Hydraulic Research (IAHR) Symposium on "Scale Effects in Porous Media" Thessaloniki, Greece, 1978.

[11] J. Bear "Hydraulics of Groundwater" McGraw-Hill Inc, 1979.

[12] R. Wiest "Flow Through Porous Media" Academic Press, 1969.